Radiomics

RADIOMICS

IMAGES ARE MORE THAN MEETS THE EYE

Ralph T.H. Leijenaar

Omslag: Schilderij door Hubertus M. Leijenaar
Druk: Datawyse | Universitaire Pers Maastricht
ISBN: 978 94 6159 781 6

Radiomics
Images are more than meets the eye

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van de Rector Magnificus Prof. dr. Rianne M. Letschert,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen,
op dinsdag 12 december 2017 om 10.00 uur

door

Ralph Theodoor Hubertina Leijenaar

Promotor

Prof. dr. Ph. Lambin

Copromotor

Dr. ir. W.J.C. van Elmpt

Dr. F.J.P. Hoebers

Beoordelingscommissie

Prof. dr. ir. W.H. Backes	(voorzitter)
Prof. dr. F.C.S. Ramaekers	
Prof. dr. M.J. Dumontier	
Prof. dr. W.J. Niessen	(Erasmus Medisch Centrum Rotterdam)
Prof. dr. M.W.M. van den Brekel	(Universiteit van Amsterdam, Nederlands Kanker Instituut)

The work presented in this thesis was made possible by the financial support of: the ERC advanced grant (ERC-ADG-2015, No. 694812 - Hypoximmuno), the QuIC-ConCePT project (IMI JU; grant No. 115151), the Dutch technology Foundation STW (grant No. 10696 DuCAT & No. P14-19 Radiomics STRaTegy), the EU 7th framework program (ARTFORCE – No. 257144, REQUITE – No. 601826), SME Phase 2 (EU proposal 673780 – RAIL), EUROSTARS (DART), the European Program H2020-2015-17 (BD2Decide - PHC30-689715 and ImmunoSABR – No. 733008), Interreg V-A Euregio Meuse-Rhine (Euradiomics), Alpe d'HuZes-KWF (DESIGN), Kankeronderzoekfonds Limburg from the Health Foundation Limburg, the Zuyderland-MAASTRO grant and the Dutch Cancer Society.

Contents

Chapter 1

General introduction and outline of the thesis

Ralph T.H. Leijenaar

INTRODUCTION

Global cancer statistics indicate that cancer is a leading cause of death worldwide, with an estimated 14.1 million new cases and 8.2 million deaths in 2012 [1, 2]. By 2030, the global cancer burden is expected to grow to 21.7 million new cancer cases and 13 million cancer deaths due to population growth and aging. Yet, this might be considerably higher due to the adoption of lifestyles that are known to increase cancer risk, including smoking, poor dietary habits and lack of physical activity [1, 2].

Up till now, treatment approaches for individual cancer patients have been based on population-based evidence, mostly from general clinical practices or clinical trials. It has become apparent that tumors and patients might be more heterogeneous than previously assumed, which calls for moving ahead from a population-based approach to precision oncology [3]. With an ever-increasing number of treatment options are becoming available, all potentially relevant available factors should therefore be taken into account in order to select the most optimal treatment [4-6].

Medical imaging is the cornerstone for the management of patients with all kinds of diseases, especially for cancer, and is used for diagnosis, treatment planning, monitoring treatment response and image-guided interventions. However, besides relatively few routine quantitative metrics, such as RECIST [7], images are mostly used in a qualitative manner in current clinical practice.

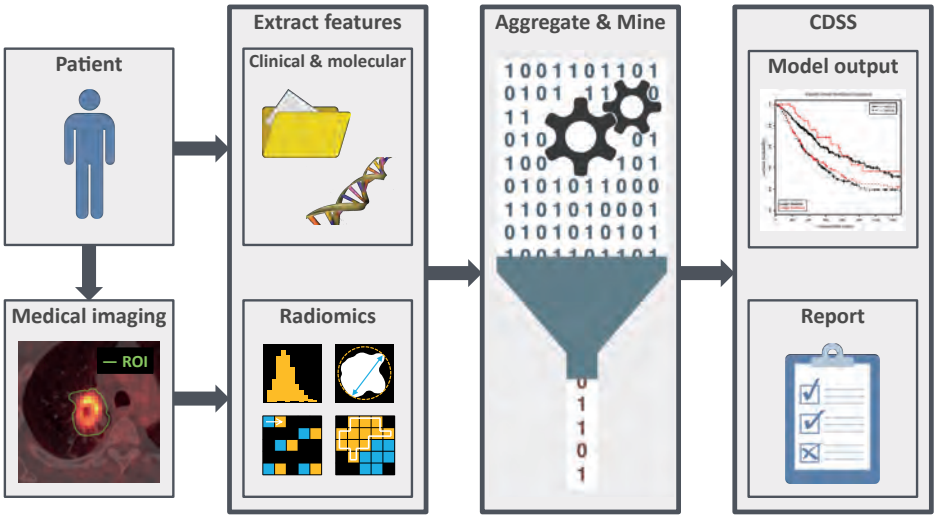


Figure 1 – Schematic example of the development of a clinical decision support system (CDSS) with radiomics. The first step for radiomics is the acquisition and segmentation (ROI: Region of Interest) of medical images. From this image data, radiomic features are extracted and stored in a database. For each patient, data from other relevant sources of information, such as clinical and molecular data, are collected and added to a database. These data sources are subsequently merged and this combined database is then analyzed to develop diagnostic, therapeutic, prognostic and predictive models.

A recent promising technique that has the potential to improve outcome for diseases for which medical imaging is used extensively, is radiomics. A concept which was first presented in a paper we published in 2012 [8]. Driven by the hypothesis that (standard of care) medical images contain quantifiable information about the underlying pathophysiology, radiomics concerns with the high-throughput mining of large amounts of quantitative features from these images, to extract knowledge [8-13].

In general, the process of radiomics consists of several steps, which are depicted in **Figure 1**. The first step is image acquisition and segmentation of regions of interest (e.g. the primary tumor volume). A large number of quantitative features is subsequently extracted from these segmented volumes. These features can broadly be divided into four groups, which quantify I) the histogram of intensity values, II) shape and size, III) texture and IV) parameters after image filtering. Using intricate modelling techniques, these features are then further analyzed for their association with a pre-defined outcome.

Given that medical imaging is an essential part of routine clinical practice, radiomics is expected to take an important place in modern clinical research [5], by cost-effectively providing complementary and interchangeable information alongside other sources, such as demographics, pathology, genomics and proteomics. The development of Clinical Decision Support Systems (CDSS), which can integrate this high dimensional patient specific information, has the potential to allow for better and more informed clinical decision making [14] (**Figure 1**). Radiomics has great potential to improve CDSS, contributing to optimizing precision oncology and a better prognosis for patients [15].

OBJECTIVES AND OUTLINE OF THE THESIS

Each step in the process of radiomics poses unique challenges and with the prospect of multicenter clinical applications, standardization and interoperability are among the biggest issues for radiomics [10-12, 16, 17]. Therefore, to facilitate further development of the field of radiomics, this thesis aims to provide deeper understanding of fundamental technical and methodological aspects. Furthermore, potential clinical applications are covered, followed by an in-depth discussion of radiomics.

Additionally, part of the work presented in this thesis consists of the development of software necessary to perform the extraction of radiomic features from medical images, in order to facilitate the research carried out in radiomics. A description of the developed software and its functionality, as used throughout the entire thesis, is provided in the addendum **Software development**.

1. Introduction

This thesis is divided into four main parts. This **Chapter 1** provides a general introduction to the thesis, which is further extended with a comprehensive Dutch review on radiomics

in Chapter 2. Chapters 3-5 are dedicated to investigating fundamental technical and methodological aspects of radiomics, in particular related to variability of radiomic features and interoperability. Chapters 6-8 focus on development and validation of potential clinical applications of radiomics in the context of lung and head and neck cancer. Chapters 9-10 provide an in-depth discussion and future prospects.

2. Technical and methodological aspects of radiomics

In Chapter 3, the aim is to investigate the stability of radiomic features. This work therefore performs an integrated stability analysis of a large number of Positron Emission Tomography (PET) derived features in non-small cell lung carcinoma (NSCLC), based on test-retest repeatability, and inter-observer reproducibility across independent manual tumor delineations of five radiation oncologists.

Chapter 4 investigates differences in methodology to calculate textural features. To quantify texture, image intensities are typically discretized into a reduced number of discrete bins. The main objective of this study is to compare two conceptually different methods for image intensity discretization for several frequently used textural features. Using a clinical case study with repeated PET imaging, this work investigates interoperability of both methods and demonstrates the effect of different methodology on the interpretation of the assessed textural features.

In Chapter 5, the focus lies on interoperability of different software implementations of radiomics, which is illustrated by the development of a clinically relevant prognostic model. This study investigates an association of post-radiochemotherapy PET radiomics with local tumor control in head and neck squamous cell carcinoma, using two different radiomics software implementations to develop two independent outcome prediction models. The reproducibility of these models is subsequently evaluated using both software implementations on an independent validation dataset.

3. Radiomics in lung and head and neck cancer

Chapter 6, describes a pivotal proof of concept study in which a prognostic radiomic signature, based on standard of care CT images, is developed and validated in over 1000 non-small cell lung cancer and head and neck cancer patients. Furthermore, an association of this signature with underlying gene-expression patterns is investigated.

In Chapter 7, the prognostic value of this radiomic signature is further validated in a large and independent cohort of oropharyngeal squamous cell carcinoma (OPSCC) patients. This study also considers how validation results are affected by the presence of CT image artifacts (e.g. streak artifacts), which are mostly caused by metallic dental fillings or other high

atomic number material implants and frequently occur on CT images of head and neck cancer patients [18].

The work presented in Chapter 8 evaluates whether human papilloma virus (HPV) status of OPSCC patients can be objectively identified by a quantitative radiomic approach, driven by the hypothesis that molecular information can be inferred from standard of care medical images. This multicenter study involves the development and validation of a CT based radiomic signature for HPV status, on a large and international collection of data from 778 OPSCC patients from four different institutions. In line with the previous chapter, the influence of CT artifacts is considered for both signature development and validation.

4. Discussion and future perspectives

Chapter 9 serves as the main and general discussion of this thesis, regarding the field of radiomics. This comprehensive review describes the process of radiomics and highlights its pitfalls, challenges and opportunities in the context of clinical decision making, with an emphasis on oncology. Besides discussing recent developments in the field, useful tools are provided to facilitate further progression and acceptance of radiomics.

Chapter 10 presents an extended discussion concerning the work presented in this thesis as well as further perspectives.

REFERENCES

[1] Stewart BW, Wild CP. World cancer report 2014. IARC.

[2] American Cancer Society. Global Cancer Facts & Figures 3rd Edition. Atlanta: American Cancer Society; 2015.

[3] Wu D, Wang DC, Cheng Y, Qian M, Zhang M, Shen Q, et al. Roles of tumor heterogeneity in the development of drug resistance: A call for precision therapy. Semin Cancer Biol 2017;42: 13-19.

[4] Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CM, et al. 'Rapid Learning health care in oncology' - an approach towards decision support systems enabling customised radiotherapy'. Radiother Oncol 2013;109: 159-164.

[5] Lambin P, Zindler J, Vanneste B, van de Voorde L, Jacobs M, Eekers D, et al. Modern clinical research: How rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine. Acta Oncol 2015;54: 1289-1300.

[6] Lambin P, Petit SF, Aerts HJWL, van Elmpt WJC, Oberije CJG, Starmans MHW, et al. The ESTRO Breur Lecture 2009. From population to voxel-based radiotherapy: Exploiting intra-tumour and intra-organ heterogeneity for advanced treatment of non-small cell lung cancer. Radiotherapy and Oncology;96: 145-152.

[7] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer;45: 228-247.

[8] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48: 441-446.

[9] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5: 4006.

[10] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol In press.

[11] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278: 563-577.

[12] Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present... any future? Eur J Nucl Med Mol Imaging 2017;44: 151-165.

[13] http://www.radiomics.world/ (accessed on: 16-08-2017)

[14] Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. Nat Rev Clin Oncol 2013;10: 27-40.

[15] Aerts HJ. The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review. JAMA Oncol 2016;2: 1636-1642.

[16] Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Scientific reports 2016;6: 23428.

[17] van Velden FH, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. Mol Imaging Biol 2016.

[18] Purohit BS, Ailianou A, Dulguerov N, Becker CD, Ratib O, Becker M. FDG-PET/CT pitfalls in oncological head and neck imaging. Insights into imaging 2014;5: 585-602.

Chapter 2

Radiomics: de toekomst in medische beeldvorming

Gepubliceerd in: **Nederlands Tijdschrift voor Oncologie.** 2017;14: 82-89.

Radiomics: de toekomst in medische beeldvorming

Ralph T.H. Leijenaar, Evelyn E.C. de Jong, Ruben T.H.M. Larue, Janna E. van Timmeren, Philippe Lambin

SAMENVATTING

Radiomics is een proces waarbij (standaard) medische beelden worden verwerkt tot kwantitatieve data, met als doel deze data–bijvoorbeeld in de vorm van diagnostische, prognostische of predictieve radiomics-modellen–te integreren in klinische keuzehulp systemen. Deze systemen bieden ondersteuning bij het maken van klinische beslissingen, wat kan bijdragen aan geoptimaliseerde, gepersonaliseerde geneeskunde en een betere prognose voor de patiënt. Radiomics is een snelgroeiend onderzoeksveld en de verwachting is dat het een steeds belangrijkere rol zal gaan spelen in de medische wereld. In dit overzicht wordt de werkwijze van radiomics toegelicht en worden recente ontwikkelingen en uitdagingen besproken. Daarnaast wordt een blik geworpen op de toekomst, met toepassingen om de gepersonaliseerde geneeskunde verder te verbeteren met radiomics.

SUMMARY

Radiomics is the high-throughput mining of quantitative image features from (standard-of-care) medical imaging for knowledge extraction. Radiomics has application within clinical decision support systems, by incorporating, for instance, diagnostic, prognostic, and predictive radiomic signatures. These decision support systems allow for better clinical decision making, which can contribute to optimizing personalized medicine and a better prognosis for patients. Radiomics is expected to have substantial implications for the medical community and is a rapidly emerging field. This review describes the process and challenges of radiomics, recent developments, and discusses future perspectives to further improve personalized medicine

INTRODUCTIE

Medische beeldvorming speelt een steeds grotere rol binnen de oncologie. Naast een diagnostische toepassing wordt het gebruikt bij het opstellen van behandelplannen en ook steeds vaker om de respons op behandeling te monitoren. Beeldvorming wordt echter in de huidige klinische praktijk vaak enkel kwalitatief gebruikt.

Radiomics, een concept voor het eerst gepresenteerd in 2012 door Lambin et al. [1], is een technologische ontwikkeling waarbij een objectieve, kwantitatieve analyse van medische beelden wordt toegepast (zie ook: www.radiomics.world). Radiomics is een proces waarbij standaard medische beelden worden verwerkt tot multidimensionale data, gedreven door de hypothese dat medische beelden kwantificeerbare informatie over de onderliggende pathofysiologie bevatten.

Radiomics heeft onder andere potentie in de gepersonaliseerde geneeskunde. Door radiomics data met gegevens van verschillende vakgebieden als genetica en bio-informatica te integreren, kunnen klinische keuzehulp systemen ontwikkeld worden (**Figuur 1**). Deze systemen zullen het mogelijk maken om betere klinische beslissingen te maken, wat kan leiden tot geoptimaliseerde, gepersonaliseerde geneeskunde en wellicht een betere prognose voor de patiënt. Radiomics is een snelgroeiend onderzoeksveld en we zien de term “radiomics” in toenemende mate terug in de literatuur (**Figuur 2**). De verwachting is dan ook dat het een belangrijke plek in modern klinisch onderzoek zal innemen [2]. In dit overzicht beschrijven we de werkwijze van radiomics en bespreken we recente ontwikkelingen, uitdagingen en onze visie voor de toekomst.

PROCES EN UITDAGINGEN

Het radiomics proces bestaat uit een aantal stappen (**Figuur 1**), achtereenvolgens zijn dat: beeldacquisitie, segmentatie van volumes, de extractie van radiomics beeldeigenschappen en het ontwikkelen en valideren van modellen.

In elke stap van het radiomics proces kunnen verschillende factoren zorgen voor variabiliteit in berekende beeldkenmerken [3, 4]. Zowel radiomics beeldkenmerken, alsmede de resultaten die verkregen zijn in (multicentrische) studies dienen één-op-één vergelijkbaar te zijn om significante en betekenisvolle conclusies te kunnen trekken. Standaardisatie is dan ook een van de grootste uitdagingen voor radiomics.

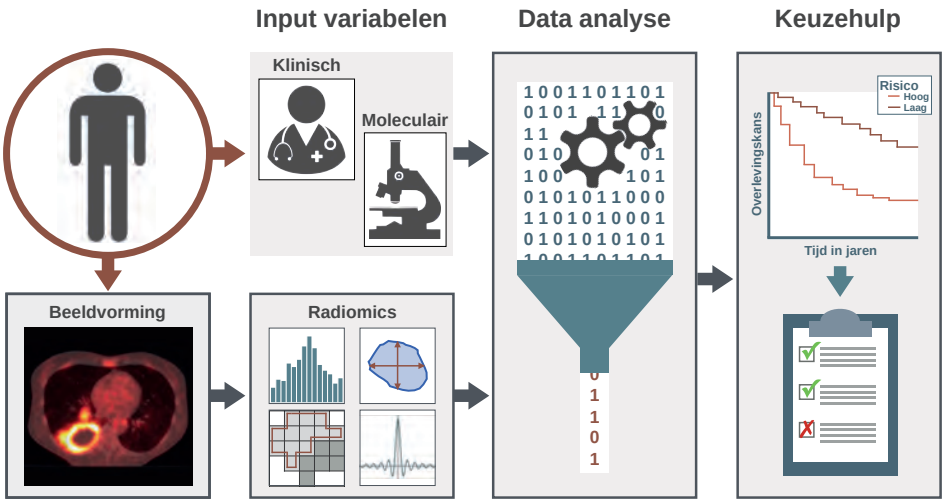
Beeldacquisitie

Radiomics begint met het verkrijgen van medische beelden met behulp van Computed Tomography (CT), Positron Emission Tomography (PET) of Magnetic Resonance Imaging (MRI). Moderne CT, PET en MR-scanners ondersteunen een heel scala aan variaties in

acquisitie- en reconstructie instellingen om subjectieve vereisten van de expert aan medische beelden te faciliteren.

Een recente studie van Mackin et al. [5] heeft aangetoond dat radiomics textuur parameters die geëxtraheerd zijn uit CT-beelden van een fantoom kunnen variëren per CT-scanner. Deze variaties kunnen zorgen voor een bias die de werkelijke onderliggende (biologische) karakteristieken maskeert. Er zijn verschillende initiatieven om beeldacquisitie en reconstructie te standaardiseren, voor zowel PET (NEDPAS [6], EANM [7]), CT (ICRU [8]) en MRI (AAPM [9]). Daarnaast zijn er internationale organisaties of consortia die beeldacquisitie verbeteren om het gebruik van kwantitatieve beeldbiomarkers mogelijk te maken, zoals het Quantitative Imaging Network (QIN) [10] en het QuIC-ConCePT project (www.quick-concept.eu) van het Innovative Medicine Initiative Joint Undertaking (IMI JU).

Naast inter-scanner variabiliteit en het gebruik van verschillende acquisitie en reconstructie parameters kan ook de veranderende anatomie van de patiënt de kwantificatie van radiomics beeldkenmerken beïnvloeden.



Figuur 1 – Overzicht van de stappen om tot een klinisch keuzehulp systeem met radiomics te komen. De eerste stap voor radiomics is de acquisitie en segmentatie van medische beelden. Uit deze beelddata worden vervolgens radiomics beeldeigenschappen geëxtraheerd en in een database opgeslagen. Van elke patiënt wordt tevens klinische en moleculaire data in een database verzameld. Deze bronnen van data worden samengevoegd en deze gecombineerde database wordt vervolgens geanalyseerd om diagnostische, prognostische en predictieve modellen te ontwikkelen.

Segmentatie van volumes

De volgende stap in het radiomics proces is het identificeren van een of meerdere (sub)volumes waaruit de radiomics features worden geëxtraheerd. Dit is veelal de primaire tumor, maar ook lymfekliermetastasen, afstandsmetastasen etc. kunnen worden gebruikt. Deze (sub-)volumes hebben vaak onduidelijke grenzen, met als gevolg dat handmatig intekenen leidt tot een significante variatie in segmentatie tussen verschillende waarnemers [11]. Dit heeft uiteraard ook invloed op de geëxtraheerde beeldeigenschappen [4, 11]. Segmentatie met behulp van (semi-)automatische methodes heeft een hogere reproduceerbaarheid [12-14] en heeft daarom de voorkeur bij radiomics studies. Er is momenteel echter nog geen semiautomatische segmentatiemethode als gouden standaard bestempeld.

Extractie van beeldeigenschappen

De meest essentiële stap van het radiomics proces is de extractie van kwantitatieve beeldeigenschappen uit de gesegmenteerde volumes. De radiomics beeldeigenschappen zijn grofweg te verdelen in vier groepen, namelijk I) intensiteit, II) vorm, III) textuur en IV) parameters na beeldtransformatie en filtering. De eerste groep, intensiteit, beschrijft het histogram van intensiteitswaarden. De tweede groep beschrijft de drie dimensionale vorm en afmetingen van het gesegmenteerde volume. Textuur, de derde groep, beschrijft intra-tumor heterogeniteit door de ruimtelijke relatie tussen beeldpunten (voxels) binnen het tumor volume te kwantificeren. De vierde groep kwantificeert intensiteit en textuur na beeldtransformatie door middel van het toepassen van filters.

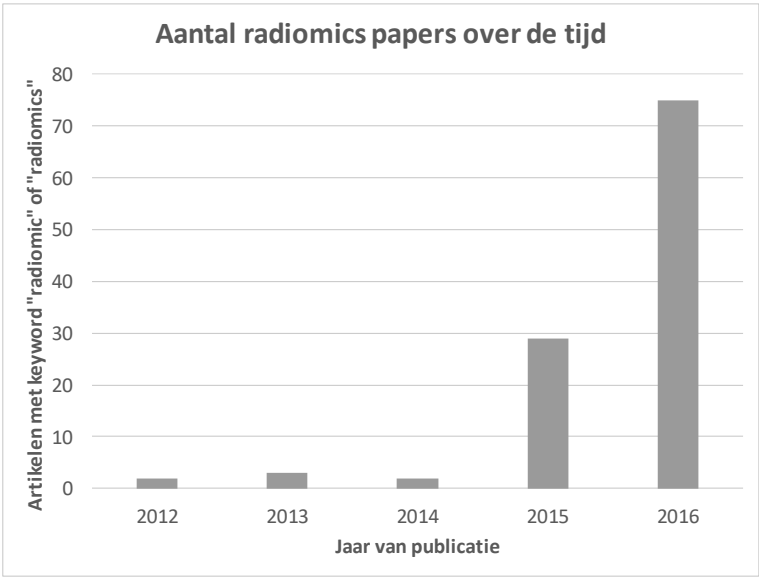
Uit recente literatuur van Hatt et al. [15] kan geconcludeerd worden dat er zowel variatie als fouten in de naamgeving, mathematische definitie en methodiek van radiomics beeldkenmerken zit, alsmede in de software implementatie van toegepaste algoritmes. Variatie in naamgeving en mathematische definities maken het moeilijk om te bepalen of verschillende radiomics studies naar dezelfde beeldeigenschappen kijken. Een recente studie door Leijenaar et al. [16] licht een van de methodologische aspecten toe die kunnen variëren tussen verschillende studies. Voorafgaand aan de analyse wordt kwantisatie van de intensiteitswaarden vaak gebruikt om de invloed van ruis te minimaliseren. De wijze waarop dit is geïmplementeerd heeft invloed op de waarden en de interpretatie van radiomics beeldkenmerken. Om radiomics beeldkenmerken te kwantificeren uit medische beelden worden software implementaties ontwikkeld, waarvan sommige als open source te downloaden zijn [17, 18]. Er zijn echter geen concrete richtlijnen voor een gestandaardiseerde implementatie van radiomics algoritmes en dit zorgt er dan ook voor dat interoperabiliteit van verschillende software niet gewaarborgd is. Deze bronnen van variatie belemmeren het trekken van conclusies uit (gepubliceerde) multicentrische radiomics studies. Een bepaald radiomics beeldkenmerk kan bijvoorbeeld in twee verschillende studies dezelfde naam en/of definitie hebben, maar verschillen in implementatie

en methodologie kunnen ervoor zorgen dat een directe vergelijking tussen beide studies niet mogelijk is. Om in de toekomst de interoperabiliteit van radiomics te kunnen waarborgen, dienen de hiervoor vermelde verschillen opgehelderd te worden.

We moeten erkennen dat er momenteel geen gouden standaard bestaat waaraan iedereen zich kan, of zelfs moet, conformeren, maar het is duidelijk dat standaardisatie van belang is voor radiomics. We raden daarom ook aan dat studies op zijn minst in detail rapporteren welke definities voor beeldkenmerken zijn gebruikt, welke methodologie is toegepast en tevens welke software implementatie is gebruikt, om verdere (externe) validatie van studieresultaten te faciliteren.

Modelontwikkeling en validatie

De laatste stap in het radiomics proces is de ontwikkeling van modellen en de validatie ervan. In het modelontwikkelingsproces is het van groot belang dat het risico op overfitten wordt geminimaliseerd. Dit resulteert namelijk in modellen die specifiek geoptimaliseerd zijn voor de data (inclusief de ruis) die gebruikt is voor modelontwikkeling. Dergelijke modellen zullen daarom slecht presteren op nieuwe data. Rangschikken van beeldeigenschappen op basis van reproduceerbaarheid, bijvoorbeeld door gebruik te maken van “test-retest” datasets, kan helpen om overfitten te voorkomen [19, 20]. Bovendien kan het aantal beeldeigenschappen worden gereduceerd groepen sterk gecorreleerde beeldeigenschappen te elimineren.



Figuur 2 – Overzicht van het toenemend aantal radiomics artikelen gepubliceerd per jaar vanaf 2012. Resultaat van een PubMed zoekopdracht naar artikelen die refereren naar “radiomics” of “radiomic”, dd. 01-12-2016.

De volgende stap in modelontwikkeling is “data mining”, oftewel het ontdekken van patronen in grote datasets. Bij data gedreven modelontwikkeling worden geen aannames gemaakt over de betekenis en het gewicht van individuele beeldeigenschappen, waarvoor “machine learning” technieken kunnen worden gebruikt (bijv. LASSO, support vector machines, Bayesian networks, etc. [21, 22]). Anderzijds, bij hypothese gedreven modelontwikkeling worden beeldeigenschappen gegroepeerd op basis van vooraf gedefinieerde informatie en klinische context. Voor beide methoden is het van belang een goed gedefinieerde uitkomst te hebben (bijv., overlevingstijd of progressievrije overleving). Idealiter wordt radiomics gecombineerd met niet-radiomics data, inclusief reeds bekende klinische voorspellers. Hierdoor kan de toegevoegde waarde van radiomics getoetst worden en kunnen tevens nieuwe modellen ontwikkeld worden die biologische, klinische en beeldinformatie bevatten, wat vervolgens kan leiden tot waardevolle diagnostische, prognostische of predictieve informatie.

Een model is echter pas van waarde wanneer het is onderworpen aan een adequate validatie. Het is daarom noodzakelijk een onafhankelijke validatie dataset ter beschikking te hebben, bij voorkeur van een ander instituut dan waar de data voor modelontwikkeling vandaan komt (externe validatie). Indien er geen validatie dataset beschikbaar is, dan dient er op zijn minst een inschatting gemaakt te worden over de modelprestaties bij nieuwe data, bijvoorbeeld door de beschikbare data op te splitsen in een deel voor modelontwikkeling en een deel voor validatie (interne validatie). Methodes als kruis-validatie of bootstrapping zijn echter een beter alternatief indien een validatie dataset ontbreekt. Tot slot is het wel cruciaal om de gebruikte statistische methodes fatsoenlijk te documenteren, zodat de validiteit van gepubliceerde modellen beoordeeld kan worden. Het TRIPOD-statement [23], bestaande uit een checklist van 22 items voor transparante documentatie van dergelijke studies, zou hiervoor een goede richtlijn kunnen zijn.

RADIOMICS, MEER DAN ALLEEN EEN CONCEPT

Dat radiomics meer is dan alleen een concept, kan afgeleid worden uit het toenemend aantal publicaties in de afgelopen jaren (**Figuur 2**), welke grotendeels zijn beschreven in (recente) overzichtspublicaties [24-31]. Er zijn radiomics studies uitgevoerd voor een groot aantal indicaties, zoals long-, hoofd-hals-, prostaat-, rectum-, slokdarmkanker en hersentumoren. In deze paragraaf zullen we enkele van deze resultaten beschrijven, welke laten zien dat radiomics een goed hulpmiddel kan zijn in de kliniek.

In 2014 heeft onze groep een op standaard planning-CT gebaseerd radiomics-profiel ontwikkeld dat in staat blijkt de prognose van de patiënt te voorspellen, voor zowel longkankerpatiënten als ook voor hoofd-halskanker patiënten [32]. Dit radiomics-profiel is gevalideerd in onafhankelijke datasets, in totaal bestaande uit meer dan 1000 patiënten. Tevens is aangetoond dat dit profiel ook prognostisch is bij aanwezige CT-artefacten (bijvoorbeeld door tandheelkundige vullingen) [33]. Een animatie die de resultaten van deze

studie samenvat is te vinden onder de volgende link: https://youtu.be/Tq980GEVP0Y. Hoewel het eerdergenoemde radiomics-profiel zowel in long als in hoofd-halskanker prognostisch is, laat een studie door Parmar et al. [34] zien dat op CT gebaseerde radiomics beeldkenmerken op specifieke wijze clusteren voor beide kankertypes, wat aantoont dat er zowel generieke als ziekte specifieke radiomics informatie te kwantificeren is.

Een recente studie door Coroller et al. [35] laat zien dat een radiomics-profiel van onafhankelijke voorspellende waarde is voor het ontwikkelen van afstandsmetastasen na radiotherapie, alsmede complementaire informatie bevat ten opzichte van klinische factoren. Dit resultaat is gevalideerd in een interne validatie dataset. Gezien er bij de behandeling van longkanker een keuze gemaakt moet worden of adjuvante chemotherapie gegeven zal worden, heeft een indicatie van het risico op afstandsmetastasen zeker klinische consequenties.

Naast CT radiomics is ook PET radiomics en MR radiomics van potentiele waarde. Zo blijkt het met behulp van MR radiomics mogelijk om prostaatkanker te onderscheiden van goedaardig prostaatweefsel en geeft het daarnaast ook informatie over de agressiviteit van de prostaatkanker [36]. Een andere studie van Teruel et al. [37] laat met behulp van één dataset zien dat voor borstkanker de respons op therapie voorspeld kan worden met behulp van radiomics, waardoor ook, indien mogelijk, een andere therapie keuze gemaakt kan worden. Een studie van Cook et al. [38] heeft laten zien dat textuur beeldeigenschappen van een 18F-FDG PET-scan respons op therapie kunnen voorspellen voor niet-kleincellige longkanker, die geassocieerd is met overleving. In deze studie hadden de textuur beeldeigenschappen zelfs een betere voorspellende waarde dan de standaard berekende PET-parameters die onder andere het metabool tumor volume vertegenwoordigen. Voor deze studie is de PET-scan gebruikt die gemaakt is voor behandeling, wat als voordeel heeft dat eventuele bijeffecten van behandeling op de FDG opname, zoals FDG opname in een door bestraling ontwikkelde longontsteking, de interpretatie van de PET-scan niet beïnvloeden [38]. Fried et al. [39] concludeerde dat het toevoegen van kwantitatieve PET-parameters de voorspellende waarde van predictiemodellen bestaande uit enkel conventionele parameters verbetert. In beide studies is de voorspellende waarde van de textuur beeldeigenschappen niet gevalideerd met behulp van externe datasets, maar in de laatstgenoemde studie is gebruik gemaakt van de kruis-validatie methode om de modellen te ontwikkelen.

Een CT radiomics studie in hepatocellulair carcinoom en een MR radiomics studie in glioblastoom waren de eersten die een relatie laten zien tussen kwantitatieve beeldvormingseigenschappen en genexpressieprofielen [40, 41]. Een studie van Panth et al. [42] heeft als “proof of principle” een causaal verband tussen radiomics en genetische veranderingen in combinatie met bestraling aangetoond.

TOEKOMSTPERSPECTIEVEN EN CONCLUSIE

Een voor de hand liggende toepassing van radiomics ligt in het gebruik ervan op meerdere complementaire standaard beeldvormingsmodaliteiten zoals (dual-energy) CT, PET en MRI, die tumor heterogeniteit elk op hun eigen wijze vastleggen. Zo kan het gebruik van PET-tracers als FMISO, FAZA of HX4 [43-46] bijvoorbeeld informatie over hypoxische heterogeniteit toevoegen. Het combineren van verschillende modaliteiten biedt mogelijkheden om tot een nog uitgebreidere karakterisering van de tumor te komen, waarbij de informatie in (standaard) beeldvorming maximaal wordt benut.

Het is alom bekend dat solide tumoren vaak heterogeen zijn. Ze bestaan uit sub-kolonies van kankercellen en veranderen in de tijd. Het nemen van biopsieën is een veelgebruikte techniek om de tumor te karakteristieken op moleculair niveau, met als beperkingen dat slechts een deel van de tumor geanalyseerd wordt en dat herhaalde biopsieën het risico voor de patiënt vergroten. Gezien het toenemende bewijs in de literatuur dat heterogeniteit implicaties heeft voor de behandeling van patiënten [47], radiomics als methode om deze tumor heterogeniteit objectief te kwantificeren. Dit wordt verder gestaafd door nieuwe inzichten, zoals dat genetische heterogeniteit een negatieve prognostische factor is voor conventionele therapie, maar juist een zeer goede prognose aanduidt bij de toepassing van immunotherapie [48]. Door het gebruik van radiomics als virtuele biopsie kunnen de beperkingen van conventionele biopsieën ondervangen worden op een niet-invasieve manier.

Een volgende stap is het introduceren van een tijdscomponent, een zogeheten ‘delta radiomics’ methode (**Figuur 3**). Voor kankerpatiënten die bestraald worden zou dit kunnen inhouden dat dagelijkse cone-beam CT-scans van een conventionele lineaire versneller (linac) [49, 50], of MRI-beelden van een MR-gestuurde linac, structureel geanalyseerd worden. De evolutie van kwantitatieve parameters die uit deze beelden worden geëxtraheerd, kan belangrijke informatie verschaffen over het effect van de therapie en de eventuele noodzaak tot tijdig aanpassen van de behandeling.

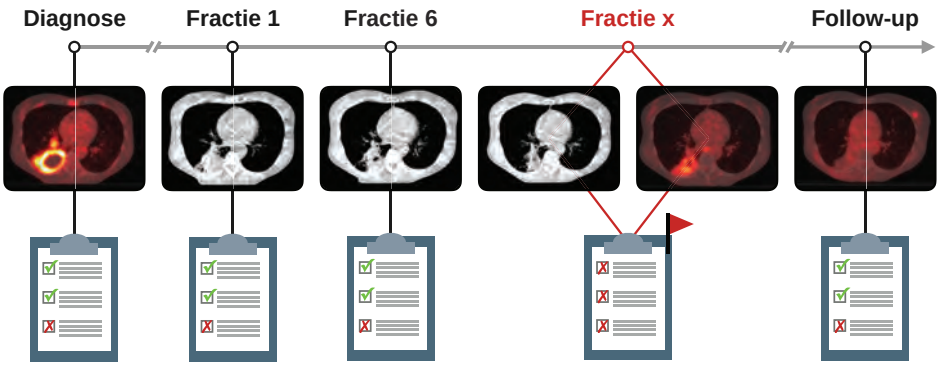
Een belangrijke ontwikkeling voor de toekomst van radiomics is het zogeheten “distributed learning”. Voor de ontwikkeling en validatie van nieuwe modellen en het verbeteren van bestaande modellen is namelijk een constant toenemende hoeveelheid patientdata nodig. Steeds meer patiëntdata wordt online beschikbaar gesteld (zie bijvoorbeeld: www.cancerdata.org), maar ten gevolge van ethische, politieke, administratieve of technische bezwaren wordt de data ook vaak niet gedeeld. Met behulp van wereldwijde “rapid learning healthcare” netwerken kan radiomics data volledig worden geïntegreerd met andere informatie over patiënten (demografisch, biologisch, etc.), behandelingen, uitkomsten en standaard beeldvorming [51, 52]. Dit biedt mogelijkheden voor distributed learning, waarmee het mogelijk is om klinische keuzehulp modellen te ontwikkelen, te verbeteren en te valideren, terwijl privacygevoelige data binnen de muren van ieders instituut blijft–enkel geaggregeerde data wordt gedeeld tussen een centrale server

en lokale dataservers. Een animatie die dit concept toelicht is te vinden onder de volgende link: https://youtu.be/ZDJFOxpwqEA.

We voorzien dat radiomics in de nabije toekomst een steeds belangrijkere rol gaat spelen. Beslishulpen en voorspellende modellen zullen steeds vaker de kennis van radiomics parameters gebruiken, die geëxtraheerd zijn uit 'rapid learning healthcare' netwerken. Daarnaast zal radiomics ook zijn weg vinden naar de radiologie, waar de beoordeling van beelden vooral gebaseerd is op een subjectieve interpretatie en intrinsieke kennis. Radiomics kan een objectieve, kwantitatieve en tevens probabilistische beoordeling bieden, i.e. "kwantitatieve radiologie" (een voorbeeld hiervan is screenen op longkanker [53]). Om deze visie te laten slagen is het echter cruciaal dat het belang van standaardisatie, automatisering en data-uitwisseling wordt erkend en dat dit, met betrokkenheid van artsen, wordt geïntegreerd in de standaard klinische praktijk.

AANWIJZINGEN VOOR DE PRAKTIJK

Radiomics is een proces waarbij standaard medische beelden (e.g. CT, PET, MR) worden verwerkt tot grote hoeveelheden kwantitatieve data.
Radiomics heeft de potentie om, door toevoegen van informatie uit standaard beeldvorming, klinische beslishulpsystemen te verbeteren
Implementatie van radiomics in de klinische praktijk is niet zonder uitdagingen. Met name standaardisatie is van groot belang.



Figuur 3 – Schematisch overzicht van een keuzehulp systeem dat gebruikt maakt van het -radiomics principe. Op basis van het radiomics-profiel dat gemaakt is van de diagnostische scan, kunnen veranderingen gedurende de behandeling worden gemonitord. Bij significante afwijkingen tijdens de radiotherapie kan er vervolgens besloten worden een geüpdatet bestralingsplan te maken op basis van een nieuwe (PET-)CT-scan.

FINANCIËLE ONDERSTEUNING

De auteurs zijn erkentelijk voor financiële ondersteuning van de ERC advanced grant (ERC-ADG-2015, n° 694812 - Hypoximmuno) en het QuIC-ConCePT project. Dit onderzoek wordt tevens ondersteund door de technologiestichting STW (grant n° 10696 DuCAT & n° P14-19 Radiomics STRaTegy), een NWO-onderdeel en het technologieprogramma van het Ministerie van Economische Zaken. De auteurs zijn ook erkentelijk voor financiële ondersteuning van het EU 7th framework program SME Phase 2 (EU proposal 673780 – RAIL) en EUROSTARS (DART).

REFERENTIES

[1] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48: 441-446.

[2] Lambin P, Zindler J, Vanneste B, van de Voorde L, Jacobs M, Eekers D, et al. Modern clinical research: How rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine. Acta Oncol 2015;54: 1289-1300.

[3] Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Scientific reports 2016;6: 23428.

[4] van Velden FH, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. Mol Imaging Biol 2016.

[5] Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest Radiol 2015;50: 757-765.

[6] Boellaard R, Oyen WJ, Hoekstra CJ, Hoekstra OS, Visser EP, Willemsen AT, et al. The Netherlands protocol for standardisation and quantification of FDG whole body PET studies in multi-centre trials. Eur. J. Nucl. Med. Mol. Imaging 2008;35: 2320-2333.

[7] Boellaard R, Delgado-Bolton R, Oyen WJ, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur. J. Nucl. Med. Mol. Imaging 2015;42: 328-354.

[8] Pahn G, Skornitzke S, Schlemmer HP, Kauczor HU, Stiller W. Toward standardized quantitative image quality (IQ) assessment in computed tomography (CT): A comprehensive framework for automated and comparative IQ analysis based on ICRU Report 87. Phys. Med. 2016;32: 104-115.

[9] Brinkmann D. WE-A-L100E-01: MR Data for Treatment Planning: Spatial Accuracy Issues, Protocol Optimization, and Applications (Preview of TG117 Report). Med. Phys. 2007;34: 2578-2578.

[10] Clarke LP, Nordstrom RJ, Zhang H, Tandon P, Zhang Y, Redmond G, et al. The Quantitative Imaging Network: NCI's Historical Perspective and Planned Goals. Transl. Oncol. 2014;7: 1-4.

[11] Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol 2013;52: 1391-1397.

[12] Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One 2014;9: e102107.

[13] van Baardwijk A, Bosmans G, Boersma L, Buijsen J, Wanders S, Hochstenbag M, et al. PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes. Int. J. Radiat. Oncol. Biol. Phys. 2007;68: 771-778.

[14] Velazquez ER, Parmar C, Jermoumi M, Mak RH, van Baardwijk A, Fennessy FM, et al. Volumetric CT-based segmentation of NSCLC using 3D-Slicer. Scientific reports 2013;3: 3529.

[15] Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present... any future? Eur J Nucl Med Mol Imaging 2017;44: 151-165.

[16] Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Scientific reports 2015;5: 11075.

[17] Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med. Phys. 2015;42: 1341-1353.

[18] Fang YH, Lin CY, Shih MJ, Wang HM, Ho TY, Liao CT, et al. Development and evaluation of an open-source software package "CGITA" for quantifying tumor heterogeneity with molecular images. Biomed Res Int 2014;2014: 248505.

[19] Balagurunathan Y, Kumar V, Gu Y, Kim J, Wang H, Liu Y, et al. Test-retest reproducibility analysis of lung CT image features. Journal of digital imaging 2014;27: 805-823.

[20] van Timmeren JE, Leijenaar RT, van Elmpt W, Wang J, Zhang Z, Dekker A, et al. Test–Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? Tomography 2016;2: 361-365.
[21] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 1996: 267-288.
[22] Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their Applications 1998;13: 18-28.
[23] Collins GS, Reitsma JB, Altman DG, Moons KM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement. Ann. Intern. Med. 2015;162: 55-63.
[24] Davnall F, Yip CS, Ljungqvist G, Selmi M, Ng F, Sanghera B, et al. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? Insights into imaging 2012;3: 573-589.
[25] O'Connor JP, Rose CJ, Waterton JC, Carano RA, Parker GJ, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. Clin Cancer Res 2015;21: 249-257.
[26] Aerts HJ. The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review. JAMA Oncol 2016.
[27] Lee G, Lee HY, Park H, Schiebler ML, van Beek EJ, Ohno Y, et al. Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art. Eur. J. Radiol. 2016.
[28] Scrivener M, de Jong EEC, van Timmeren JE, Pieters T, Ghaye B, Geets X. Radiomics applied to lung cancer: a review. Translational Cancer Research 2016;5: 398-409.
[29] Larue RT, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. Br J Radiol 2017;90: 20160665.
[30] van Rossum PS, Xu C, Fried DV, Goense L, Court LE, Lin SH. The emerging field of radiomics in esophageal cancer: current evidence and future potential. Translational Cancer Research 2016;5: 410-423.
[31] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol In press.
[32] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5: 4006.
[33] Leijenaar RT, Carvalho S, Hoebers FJ, Aerts HJ, van Elmpt WJ, Huang SH, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta Oncol 2015;54: 1423-1429.
[34] Parmar C, Leijenaar RT, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. Scientific reports 2015;5: 11044.
[35] Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RT, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 2015;114: 345-350.
[36] Wibmer A, Hricak H, Gondo T, Matsumoto K, Veeraraghavan H, Fehr D, et al. Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. Eur. Radiol. 2015;25: 2840-2850.
[37] Teruel JR, Heldahl MG, Goa PE, Pickles M, Lundgren S, Bathen TF, et al. Dynamic contrast-enhanced MRI texture analysis for pretreatment prediction of clinical and pathological response to neoadjuvant chemotherapy in patients with locally advanced breast cancer. NMR Biomed. 2014;27: 887-896.
[38] Cook GJ, Yip C, Siddique M, Goh V, Chicklore S, Roy A, et al. Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy? J Nucl Med 2013;54: 19-26.
[39] Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, et al. Stage III Non-Small Cell Lung Cancer: Prognostic Value of FDG PET Quantitative Imaging Features Combined with Clinical Prognostic Factors. Radiology 2016;278: 214-222.
[40] Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nat. Biotechnol. 2007;25: 675-680.
[41] Diehn M, Nardini C, Wang DS, McGovern S, Jayaraman M, Liang Y, et al. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. Proc. Natl. Acad. Sci. U. S. A. 2008;105: 5213-5218.

[42] Panth KM, Leijenaar RT, Carvalho S, Lieuwes NG, Yaromina A, Dubois L, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. Radiother Oncol 2015;116: 462-466.
[43] Peeters SG, Zegers CM, Lieuwes NG, van Elmpt W, Eriksson J, van Dongen GA, et al. A comparative study of the hypoxia PET tracers [(1)(8)F]HX4, [(1)(8)F]FAZA, and [(1)(8)F]FMISO in a preclinical tumor model. Int. J. Radiat. Oncol. Biol. Phys. 2015;91: 351-359.
[44] Zegers CM, van Elmpt W, Reymen B, Even AJ, Troost EG, Ollers MC, et al. In vivo quantification of hypoxic and metabolic status of NSCLC tumors using [18F]HX4 and [18F]FDG-PET/CT imaging. Clin Cancer Res 2014;20: 6389-6397.
[45] Zegers CM, van Elmpt W, Szardenings K, Kolb H, Waxman A, Subramaniam RM, et al. Repeatability of hypoxia PET imaging using [(1)(8)F]HX4 in lung and head and neck cancer patients: a prospective multicenter trial. Eur. J. Nucl. Med. Mol. Imaging 2015;42: 1840-1849.
[46] Zegers CM, van Elmpt W, Wierts R, Reymen B, Sharifi H, Ollers MC, et al. Hypoxia imaging with [(1)(8)F]HX4 PET in NSCLC patients: defining optimal imaging parameters. Radiother. Oncol. 2013;109: 58-64.
[47] Sun XX, Yu Q. Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. Acta Pharmacol. Sin. 2015;36: 1219-1227.
[48] Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 2015;348: 124-128.
[49] Fave X, Mackin D, Yang J, Zhang J, Fried D, Balter P, et al. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? Med. Phys. 2015;42: 6784.
[50] Van Timmeren JE, Leijenaar RTH, Van Elmpt W, Lambin P. PO-0922: Are planning CT radiomics and cone-beam CT radiomics interchangeable? Radiother. Oncol. 2016;119: S446-S447.
[51] van Timmeren JE, Leijenaar RTH, van Elmpt W, Reymen B, Oberije C, Monshouwer R, et al. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiother Oncol 2017.
[52] Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CM, et al. 'Rapid Learning health care in oncology' - an approach towards decision support systems enabling customised radiotherapy'. Radiother Oncol 2013;109: 159-164.
[53] Hawkins S, Wang H, Liu Y, Garcia A, Stringfield O, Krewer H, et al. Predicting Malignant Nodules from Screening CT Scans. J. Thorac. Oncol. 2016;11: 2120-2128.

Chapter 3

Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and inter-observer variability

Published in: Acta Oncologica. 2013;52: 1391-1397.

Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and inter-observer variability

Ralph T.H. Leijenaar*, Sara Carvalho*, Emmanuel Rios Velazquez,
Wouter J.C. van Elmpt, Chintan Parmar, Otto S. Hoekstra, Corneline J. Hoekstra,
Ronald Boellard, André L.A.J. Dekker, Robert J. Gillies, Hugo J.W.L. Aerts,
Philippe Lambin

* These authors contributed equally to this work

ABSTRACT

Background

Besides basic measurements as maximum standardized uptake value $(SUV)_{max}$ or SUV_{mean} derived from ^{18}F-FDG positron emission tomography (PET) scans, more advanced quantitative imaging features (i.e. "Radiomics" features) are increasingly investigated for treatment monitoring, outcome prediction, or as potential biomarkers. With these prospected applications of Radiomics features, it is a requisite that they provide robust and reliable measurements. The aim of our study was therefore to perform an integrated stability analysis of a large number of PET-derived features in non-small cell lung carcinoma (NSCLC), based on both a test-retest and an inter-observer setup.

Methods

Eleven NSCLC patients were included in the test-retest cohort. Patients underwent repeated PET imaging within a one-day interval, before any treatment was delivered. Lesions were delineated by applying a threshold of 50% of the maximum uptake value within the tumour. Twenty-three NSCLC patients were included in the inter-observer cohort. Patients underwent a diagnostic whole-body PET-computed tomography (CT). Lesions were manually delineated based on fused PET-CT, using a standardized clinical delineation protocol. Delineation was performed independently by five observers, blinded to each other. Fifteen first order statistics, 39 descriptors of intensity volume histograms, eight geometric features and 44 textural features were extracted. For every feature, test-retest and inter-observer stability was assessed with the intra-class correlation coefficient (ICC) and the coefficient of variability, normalized to mean and range. Similarity between test-retest and inter-observer stability rankings of features was assessed with Spearman's rank correlation coefficient.

Results

Results showed that the majority of assessed features had both a high test-retest (71%) and inter-observer (91%) stability in terms of their ICC. Overall, features more stable in repeated PET imaging were also found to be more robust against inter-observer variability.

Conclusion

Results suggest that further research of quantitative imaging features is warranted with respect to more advanced applications of PET imaging as being used for treatment monitoring, outcome prediction or imaging biomarkers.

INTRODUCTION

Positron emission tomography (PET) has been shown to be a valuable tool for the detection and staging of lung cancer [1]. In recent years also PET imaging has also been increasingly used for treatment planning [2] and response monitoring in radiotherapy [3]. The most widely used tracer in oncological PET imaging is the glucose analog ^{18}F-Fluoro-2-Deoxy-D-glucose (FDG), commonly quantified by the standardized uptake value (SUV) [4]. Previous research provides evidence of basic and easily derived pre-treatment PET measurements, such as the maximum (SUV_{max}) or mean SUV (SUV_{mean}), being predictors for treatment outcome in NSCLC [5-7]. Besides these basic measurements, more advanced quantitative imaging features are increasingly investigated for treatment monitoring and outcome prediction in lung and other cancer sites [8-10], or as potential imaging biomarkers [11].

The use of basic and more advanced descriptors derived from PET imaging is within the scope of "Radiomics" [12-14]: a high throughput approach to extract and mine a large number of quantitative features from medical images, where it is hypothesized that it will improve tumour characterization and treatment outcome prediction. However, with the prospect of using these Radiomics features for future prognostic and predictive models, knowledge about their reliability and variability is needed. A few recent studies have investigated these aspects of FDG-PET derived parameters in different cancer sites, including the test-retest stability of basic SUV measurements [15], test-retest stability of a number of basic and textural features [16], or the variability of textural features due to image acquisition and reconstruction parameters [17]. However, to our knowledge no previous study has performed an integrated stability analysis of a large number of PET features in NSCLC, based on both a test-retest and an inter-observer setup. Therefore, the aim our study is to independently examine the feature's test-retest reliability and inter-observer stability between multiple manual tumour delineations. Moreover, we aim to combine the information obtained from both analyses to assess if imaging features that are more stable in repeated PET imaging are also more robust against inter-observer variability. Based on literature research, we strived to include a broad collection of PET based imaging features used in the context of predictive and/or prognostic modelling in cancer, to provide a comprehensive overview.

MATERIALS AND METHODS

This study includes two separate patient cohorts in order to assess both the test-retest and inter-observer variability of a large number of quantitative imaging features. All patients signed an informed consent form in accordance with approval by the institutional review board. A schematic representation of the workflow applied in our study is depicted in **Figure 1**.

Test-retest cohort

Eleven patients with histology- or cytology-diagnosed non–small cell lung cancer (NSCLC) were included in this patient cohort, as described in [18]. Patients underwent two baseline ^{18}F-FDG-PET scans within a one-day interval, before any treatment was delivered. PET images were acquired on an ECAT EXACT HR1 scanner (Siemens/CTI) and iteratively reconstructed using normalization- and attenuation-weighted ordered-subset expectation maximization with two iterations and sixteen subsets (OSEM 2i16s). All images had an in-plane resolution of 5.15×5.15 mm/pixel and a slice thickness of 2.43 mm. Further patient and imaging details are described by Frings et al [18]. Lesions with adequate uptake were first identified and subsequently delineated by applying a threshold of 50% of the maximum uptake value within the tumour [19], using a semiautomatic delineation tool [18] (**Figure 2**).

Inter-observer cohort

Twenty-three patients with histologically proven NSCLC were included in this patient cohort, as described previously in [20]. Patients underwent a diagnostic whole-body PET-CT scan acquired on a SOMATOM Sensation 16 with an ECAT ACCEL PET scanner (Siemens, Erlangen, Germany). PET images were iteratively reconstructed using normalization- and attenuation-weighted OSEM 4i8s. Images had an in-plane resolution of 5.31×5.31 mm/pixel and a 5 mm slice thickness. Primary tumours and involved lymph nodes were identified and manually delineated based on fused PET-CT images, using a standardized clinical delineation protocol. Delineation of the lesions was performed independently by five observers and all observers were blinded to the contours delineated by the others (Figure 2.C-D). Manual delineations were performed on XiO/Focal (Computer Medical System, Inc., St. Louis, MO). For further details on the patient cohort, imaging and delineation, we refer to the publication of van Baardwijk et al [20].

Image processing and feature extraction

All image analysis was performed in Matlab R2012b (The Mathworks, Natick, MA) using an adapted version of CERR (the Computational Environment for Radiotherapy Research) [21] extended with in-house developed Radiomics image analysis software to extract imaging features. PET images and delineated VOIs were first imported into CERR, where the image intensities were normalized to SUV [4]. First order statistics consisted of basic SUV measurements and features describing histogram of voxel intensity values contained within the VOI.

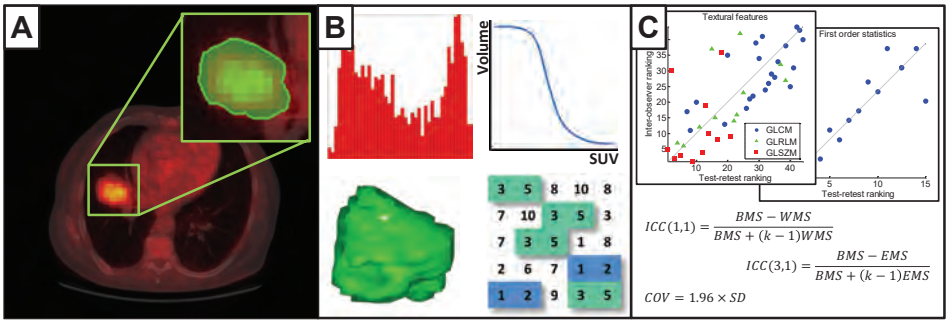


Figure 1 – Schematic of the workflow applied in our study: (A) Acquisition of PET images (fused CT for illustrative purposes), followed by tumour delineation; (B) Extraction of Radiomics features from the defined volume of interest; (C) Test-retest and inter-observer stability analysis.

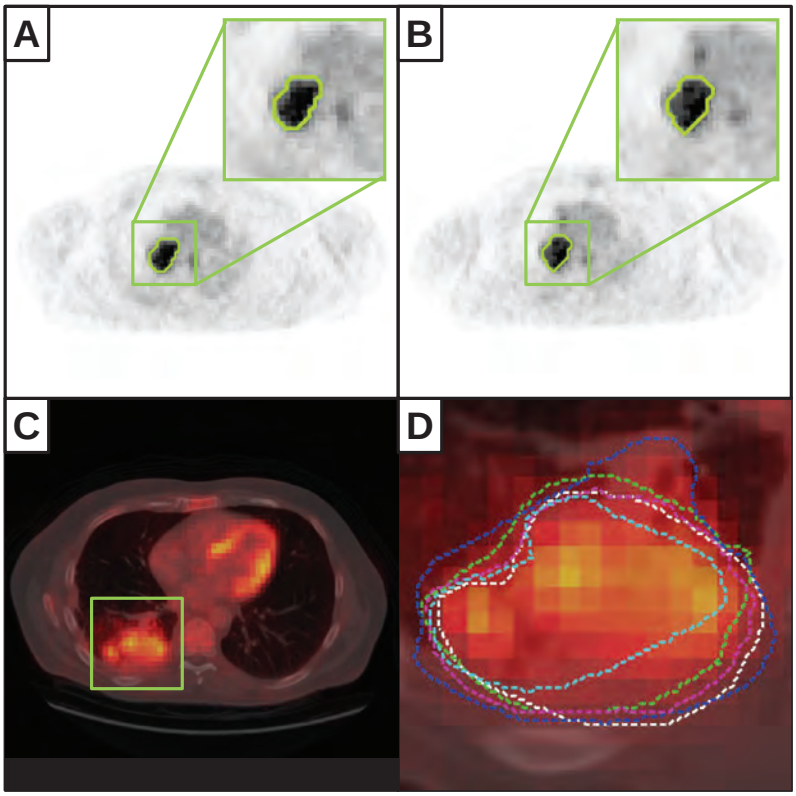


Figure 2 – (A and B) Representative images of repeated imaging of a patient from the test-retest cohort, with the 50% SUV_{max} tumour delineation shown outlined in green, for respectively the first and second baseline PET scan. (C) Representative image of a patient from the inter-observer cohort, where the lesion area is outlined with the green square (fused CT for illustrative purposes); (D) Enlargement of the lesion area with in different colours the five independent tumour delineations by multiple observers.

A set of metrics was derived from intensity volume histogram (IVH) representations [10], which summarize the complex three-dimensional data contained in the image into a single curve, allowing for a simplified interpretation. Three IVH definitions were considered: the relative volume as a function of the relative intensity (RVRIx), the absolute volume as a function of the relative intensity (AVRIx) and the intensity threshold as a function of the relative volume having a maximum intensity lower than the threshold (AIRVx). Relative steps in volume and intensity (x) were taken in 10% increments, from 10%-90%. Furthermore, three differential IVH metrics were considered: RVRIx-RVRI(100-x), AVRIx-AVRI(100-x), and AIRVx-AIRV(100-x).

Geometric features were calculated, describing the three-dimensional shape and size of the lesions. Textural features describing patterns or spatial distribution of voxel intensities, were calculated from respectively grey level co-occurrence (GLCM) [22], grey level run-length (GLRLM) [23] and grey level size-zone texture matrices (GLSZM) [9]. Determining texture matrix representations requires the voxel intensity values within the VOI to be discretized. Voxel intensities were therefore resampled into equally spaced bins using a bin-width of 0.5 units SUV. This discretization step not only reduces image noise, but also normalizes intensities across all patients, allowing for a direct comparison of all calculated textural features between patients. Texture matrices were determined considering 26-connected voxels (i.e. voxels were considered to be neighbours in all 13 directions in three dimensions) and a distance of one voxel between consecutive voxels was set for co-occurrence and grey level run-length matrices. Features derived from co-occurrence and grey level run-length matrices were calculated by averaging their value over all 13 considered directions in three dimensions.

Overall, the extracted imaging features comprised 15 first order statistics, 39 descriptors of intensity volume histograms, 8 geometric features and 44 textural features. Mathematical definitions, if applicable, for features assessed in our study can be found in Supplementary Appendix A, to be found online at http://informahealthcare.com/doi/abs/10.3109/0284186X.2013.812798

Statistical Analysis

The intra-class correlation coefficient (ICC) [24] was calculated to provide an indication of both the test-retest and inter-observer reliability of feature measurements. The ICC is a statistical measure between 0 and 1, where 0 indicates no and 1 indicates perfect reliability. To determine the ICC, variance estimates were obtained through partitioning the total variance by means of non-parametric analysis of variance (ANOVA) by ranks. To assess test-retest reliability of imaging features, we used the definition of ICC(1,1), given by:

$$ICC(1,1) = \frac{BMS - WMS}{BMS + (k-1)WMS}$$

Where BMS and WMS are respectively the between-subjects and within-subjects mean squares, obtained by Kruskal-Wallis one-way ANOVA, and k is the number of repeated measurements (i.e. PET scans). Inter-observer stability was determined with the definition of ICC(3,1), with the form:

$$ICC(3,1) = \frac{BMS - EMS}{BMS + (k-1)EMS}$$

Where BMS and EMS are the between-subjects and residual mean squares acquired from Friedman's two-way ANOVA, and k is the number of observers (i.e. delineators). Absolute variability was estimated as the coefficient of variability (COV), defined as the value below which the difference between two measurements will be with 95% probability [25]:

$$COV = 1{,}96 \times SD$$

Where SD is the standard deviation for single differences on different subjects (i.e. lesions). To provide a basis for evaluating the magnitude of the test-retest and inter-observer COV values, we normalized them to a percentage of the mean feature value ($COV_{\%mean}$) as well as the range of feature values (2.5 – 97.5 percentile; $COV_{\%range}$) over all included lesions. To assess the similarity of the test-retest and inter-observer stability rankings of features we ranked them, per feature group, in terms of their ICC. The similarity of feature rankings was determined with Spearman's rank correlation coefficient (ρ_S).

All statistical analysis was performed in Matlab R2012b (The Mathworks, Natick, MA).

RESULTS

Lesion identification and delineation resulted in a total number of 18 lesions to be included for the test-retest analysis and respectively 27 lesions for the inter-observer analysis. Test-retest and inter-observer ICC, $COV_{\%mean}$ and $COV_{\%range}$ values are summarized per feature group in respectively **Table 1** and **Table 2**, where we classified features into three groups, as having a high (ICC≥0.8), medium (0.8>ICC≥0.5), or low (ICC<0.5) stability.

AVRIx and RVRIx for x≤50% were excluded from test-retest analysis, since they represent the entire (relative) tumour volume and therefore provide no additional information on test-retest variability. In summary, 71% of all assessed features had high, 18% a medium and 11% a low stability in terms of their test-retest ICC. We found a high inter-observer stability for 91% of imaging features, whereas 8% and 1% of the features had a medium or respectively low stability. As expected, SUV_{max} and SUV_{peak} showed perfect inter-observer stability (ICC=1). Because of the same reasoning outlined above, we also excluded RVRIx and AVRIx for x≤50% from the comparative analysis. Scatter plots of stability rankings for every feature group are depicted in **Figure 3.A-D**. Considering all features, we observed a good overall similarity in feature stability rankings in terms of test-retest and inter-observer ICCs (ρ_S =0.67, p<0.01). Comparing stability rankings per feature

group, we found a high similarity for both the first order statistics (ρ_S=0.88, p<0.01) and the textural features (ρ_S=0.72, p<0.01).

As can be observed from **Figure 3.D**, features based on GLSZM have the overall lowest ranks in both analyses, indicating these features have the highest variability amongst all textural features. For the IVH features the observed similarity was more moderate (ρ_S=0.57, p<0.01). Comparing the rankings for the geometric features resulted in a non-significant ρ_s of 0.66 (p=0.09). However, from **Figure 3.C**, a positive trend in similarity can be observed. Overall, these results show that features that are more stable in repeated PET imaging are also more robust against inter-observer variability.

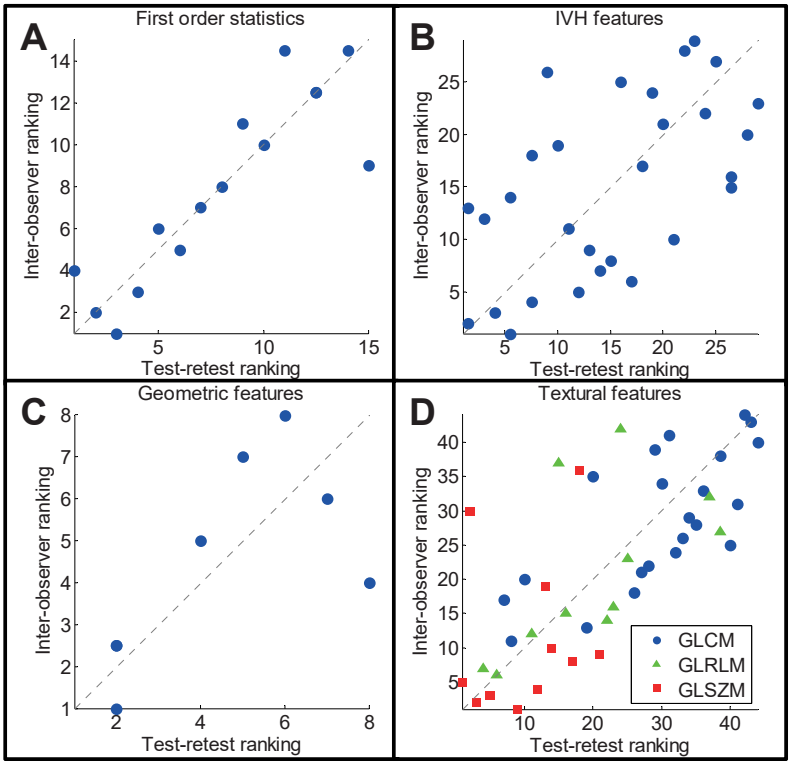


Figure 3 – Scatter plots of stability rankings of test-retest versus inter-observer intra-class correlation coefficients. Plotted diagonal illustrates perfect correlation. (A) First order statistics (ρ_S=0.877, p<<0.001$\rho_s = 0.877, p \ll 0.001$). (B) Intensity volume histogram features (ρ_S=0.572, p<<0.001$\rho_s = 0.572, p = 0.001$). (C) Geometric features (ρ_S=0.663, p<<0.001$\rho_s = 0.663, p = 0.086$). (D) Textural features (ρ_S=0.719, p<<0.001$\rho_s = 0.719, p \ll 0.001$), with GLCM features in blue circles, GLRLM features in green triangles and GLSZM features in red squares.

Table 1 – Results for the test-retest analysis, showing ICC, $COV_{\%mean}$ and $COV_{\%range}$ ranges, as well as the number of features per feature group and per class, defined as high (ICC≥0.8), medium (0.8>ICC≥0.5), or low (ICC<0.5) stability. Median values of ICC, $COV_{\%mean}$ and $COV_{\%range}$ ranges are shown within brackets.

Stability class	N	ICC	$COV_{\%mean}$ (%)	$COV_{\%range}$ (%)
First order statistics				
High stability	13	0.81 - 0.96 (0.92)	17.27 - 86.29 (23.45)	12.22 - 35.36 (14.67)
Medium stability	0	-	-	-
Low stability	2	0.27 - 0.28 (0.27)	57.29 - 110.48 (83.89)	55.61 - 60.58 (58.09)
IVH features				
High stability	18	0.80 - 0.94 (0.86)	17.09 - 44.07 (29.19)	3.39 - 23.78 (14.82)
Medium stability	3	0.61 - 0.78 (0.77)	37.26 - 105.65 (50.40)	6.03 - 28.04 (20.33)
Low stability	8	0.00 - 0.48 (0.27)	7.68 - 99.25 (46.30)	46.82 - 68.00 (60.54)
Geometric features				
High stability	8	0.81 - 0.88 (0.83)	12.25 - 37.61 (29.53)	3.80 - 31.58 (18.79)
Medium stability	0	-	-	-
Low stability	0	-	-	-
Textural features				
High stability	29	0.81 - 0.93 (0.89)	2.76 - 166.45 (36.90)	5.94 - 36.42 (19.25)
Medium stability	14	0.54 - 0.79 (0.64)	3.57 - 465.37 (75.93)	4.32 - 54.56 (33.96)
Low stability	1	0.35	84.19	53.59

Table 2 – Results for the inter-observer analysis, showing ICC, $COV_{\%mean}$ and $COV_{\%range}$ ranges, as well as the number of features per feature group and per class, defined as high (ICC≥0.8), medium (0.8>ICC≥0.5), or low (ICC<0.5) stability. Median values of ICC, $COV_{\%mean}$ and $COV_{\%range}$ ranges are shown within brackets.

Stability class	N	ICC	$COV_{\%mean}$ (%)	$COV_{\%range}$ (%)
First order statistics				
High stability	14	0.87 - 1.00 (0.98)	2.07 - 58.17 (15.25)	1.20 - 22.75 (7.39)
Medium stability	1	0.79	65.81	41.21
Low stability	0	-	-	-
IVH features				
High stability	34	0.82 - 1.00 (0.97)	5.60 - 131.45 (28.57)	1.23 - 52.15 (10.70)
Medium stability	5	0.63 - 0.77 (0.72)	4.53 - 39.04 (21.74)	38.72 - 57.65 (51.14)
Low stability	0	-	-	-
Geometric features				
High stability	8	0.80 - 0.98 (0.97)	11.63 - 48.47 (26.79)	9.60 - 31.31 (19.20)
Medium stability	0	-	-	-
Low stability	0	-	-	-
Textural features				
High stability	39	0.80 - 0.99 (0.95)	1.20 - 257.20 (28.61)	5.34 - 40.03 (13.19)
Medium stability	3	0.50 - 0.77 (0.75)	44.46 - 128.87 (104.07)	12.38 - 51.25 (30.16)
Low stability	2	0.17 - 0.19 (0.18)	156.41 - 192.86 (174.63)	57.96 - 76.36 (67.16)

DISCUSSION

Increased investigation of quantitative imaging features to monitor response to treatment, treatment outcome or as potential imaging biomarkers, raised the requisite to validate their accuracy, robustness and stability. We first independently investigated the stability of imaging features in both a test- retest and intra-observer setting and subsequently performed an integrated analysis. Our results indicated high ICC values and high stability for the majority of assessed PET image features in both the test-retest (71%) and inter-observer analysis (91%). Furthermore, we found that features that were more stable in repeated imaging were also more robust against multiple tumour delineations. These results suggest that, even though there are different sources of feature variability, one can define a set of features being overall most reliable.

We focused our results mainly on the ICC. Being a dimensionless statistic, the ICC is useful when comparing the stability of measures with different units, as is the case with the PET imaging features assessed in this study. We chose arbitrary ICC thresholds to define high, medium and low stability. There is however no consensus how high the ICC should be to for a measure to be considered to have an acceptably high reliability, since the ICC is a relative measure determined from the between and within subject (i.e. lesion) variance, which makes it a sample specific measure. This implies that ICC values obtained from our test-retest analysis were not directly comparable to those from the inter-observer analysis, since they were independently obtained from two different patient cohorts (i.e. different lesions and differences in image acquisition and reconstruction). To overcome this limitation, we ranked features according to their ICC, allowing us to compare stability rankings of features between the two analyses.

In the inter-observer analysis, SUV_{max} and SUV_{peak} both had an ICC of 1, indicating perfect stability. However, we did observe a small COV for these features, which was unexpected. A detailed look into all delineations revealed that for only one lesion, one delineator did not include the maximum uptake voxel in the delineated tumour region. Tixier et al. [16] studied the reliability of a number of basic and textural FDG-PET features in a test-retest setting in oesophageal cancer. Although the results presented in that study are not directly comparable to our test-retest results, it can be observed that textural features based on grey-level size-zone matrix representations appear to be the least stable ones, which is also supported by our test-retest, inter-observer and integrated analysis.

While the ICC is a useful tool in assessing the reliability of feature measurements, it is not directly related to a feature's clinical usefulness. For a more complete picture, one would like to know if the inter patient variability or respectively the change in feature values between a reference time point (e.g. pre-treatment) and a point of interest (e.g. during or post treatment) is large enough to be considered useful. To assess this aspect of feature variability, a measure besides the ICC is necessary that provides information on the variability in terms of the feature's unit of measurement. In our study, we therefore estimated both the test-retest and inter-observer coefficient of variability for every

feature and normalized them to a percentage of the mean feature value as well as the range, to provide easy to interpret values regarding the magnitude of the COV. The larger the COV is compared to inter patient variability or changes in feature values, the less likely it is that the feature under consideration is a useful predictor or biomarker. One has to note however, that like the ICC, COV values are sample specific estimates and typical feature values (i.e. mean and range) are likely to be different when considering different patient populations. Furthermore, the level of variation of a feature that is considered acceptable depends on its intended purpose.

A limitation of our study is the small number of patients in both cohorts. Although a broad range of tumour sizes and levels of tracer uptake were included, external validation is needed to assess if our results are representative for NSCLC patients in general. Besides feature variability due to repeated imaging and inconsistency between multiple manual tumour delineations, there are more sources of variability that can be taken into consideration. Galavis et al. [17] pointed out that quantitative imaging features are also subject to vary due to different acquisition modes and reconstruction parameters. Also, the level of image discretization has been shown to impact the variability of certain textural features, as demonstrated by Tixier et at. [16]. Taking these sources of variability into account, it is evident that standardization is desirable with the prospect of FDG-PET Radiomics features for treatment monitoring, outcome prediction or imaging biomarkers.

CONCLUSION

The aim of this study was to perform an integrated stability analysis of PET Radiomics features obtained from FDG-PET imaging in NSCLC. Our results showed that the majority of assessed features had both a high test-retest (71%) as well as inter-observer stability (91%) in terms of their intra-class correlation coefficient. Furthermore, it was observed that features more stable in repeated PET imaging were in general also more robust against inter-observer variability. Results suggest that further research of quantitative imaging features is warranted with respect to more advanced applications of PET imaging as being used for treatment monitoring, outcome prediction or imaging biomarkers.

ACKNOWLEDGEMENTS

Authors acknowledge the QuIC-ConCePT project, which is partly funded by EFPIA companies and the Innovative Medicine Initiative Joint Undertaking (IMI JU) under Grant Agreement No. 115151. Authors also acknowledge financial support from the National Institute of Health (NIH-USA U01 CA 143062-01, Radiomics of NSCLC), the CTMM framework (AIRFORCE project, grant 030-103), EU 6th and 7th framework program (EUROXY, METOXIA, EURECA, ARTFORCE), euroCAT (IVA Interreg - www.eurocat.info), Kankeronderzoekfonds Limburg from the Health Foundation Limburg and the Dutch Cancer Society (KWF UM 2011-5020, KWF UM 2009-4454).

REFERENCES

[1] Lin P, Koh ES, Lin M, Vinod SK, Ho-Shon I, Yap J, et al. Diagnostic and staging impact of radiotherapy planning FDG-PET-CT in non-small-cell lung cancer. Radiother Oncol 2011;101: 284-290.

[2] De Ruysscher D, Nestle U, Jeraj R, Macmanus M. PET scans in radiotherapy planning of lung cancer. Lung Cancer 2012;75: 141-145.

[3] Van Elmpt W, Pottgen C, De Ruysscher D. Therapy response assessment in radiotherapy of lung cancer. Q J Nucl Med Mol Imaging 2011;55: 648-654.

[4] Thie J. Understanding the standardized uptake value, its methods, and implications for usage. Journal of nuclear medicine 2004;45: 1431-1434.

[5] van Elmpt W, Ollers M, Dingemans AM, Lambin P, De Ruysscher D. Response assessment using 18F-FDG PET early in the course of radiotherapy correlates with survival in advanced-stage non-small cell lung cancer. J Nucl Med 2012;53: 1514-1520.

[6] Takeda A, Yokosuka N, Ohashi T, Kunieda E, Fujii H, Aoki Y, et al. The maximum standardized uptake value (SUVmax) on FDG-PET is a strong predictor of local recurrence for localized non-small-cell lung cancer after stereotactic body radiotherapy (SBRT). Radiother Oncol 2011;101: 291-297.

[7] Velazquez ER, Aerts HJ, Oberije C, De Ruysscher D, Lambin P. Prediction of residual metabolic activity after treatment in NSCLC patients. Acta Oncol 2010;49: 1033-1039.

[8] Vaidya M, Creach KM, Frye J, Dehdashti F, Bradley JD, El Naqa I. Combined PET/CT image characteristics for radiotherapy tumour response in lung cancer. Radiother Oncol 2012;102: 239-245.

[9] Tixier F, Le Rest CC, Hatt M, Albarghach N, Pradier O, Metges JP, et al. Intratumour heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med 2011;52: 369-378.

[10] El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit 2009;42: 1162-1171.

[11] Buckler AJ, Bresolin L, Dunnick NR, Sullivan DC. Quantitative imaging test approval and biomarker qualification: interrelated but distinct activities. Radiology 2011;259: 875-884.

[12] Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. Nat Rev Clin Oncol 2013;10: 27-40.

[13] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48: 441-446.

[14] Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging 2012;30: 1234-1248.

[15] de Langen AJ, Vincent A, Velasquez LM, van Tinteren H, Boellaard R, Shankar LK, et al. Repeatability of 18F-FDG uptake measurements in tumours: a metaanalysis. J Nucl Med 2012;53: 701-708.

[16] Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumour uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. J Nucl Med 2012;53: 693-700.

[17] Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol 2010;49: 1012-1016.

[18] Frings V, de Langen AJ, Smit EF, van Velden FH, Hoekstra OS, van Tinteren H, et al. Repeatability of metabolically active volume measurements with 18F-FDG and 18F-FLT PET in non-small cell lung cancer. J Nucl Med 2010;51: 1870-1877.

[19] Cheebsumon P, Boellaard R, de Ruysscher D, van Elmpt W, van Baardwijk A, Yaqub M, et al. Assessment of tumour size in PET/CT lung cancer studies: PET- and CT-based methods compared to pathology. EJNMMI Res 2012;2: 56.

[20] van Baardwijk A, Bosmans G, Boersma L, Buijsen J, Wanders S, Hochstenbag M, et al. PET-CT-based autocontouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumour and involved nodal volumes. Int J Radiat Oncol Biol Phys 2007;68: 771-778.

[21] Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Med Phys 2003;30: 979-985.
[22] Haralick RM, Shanmugam K, Dinstein I. Textural Features of Image Classification. IEEE T Syst Man Cyb 1973;SMC-3: 610-621.
[23] Galloway M. Texture analysis using gray level run lengths. Comput Vision Graph 1975;4: 172-179.
[24] Shrout PE, Fleiss JL. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychol Bull 1979;86: 420-428.
[25] Bland J, Altman D. Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat 2007;17: 571-582.

Chapter 4

The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis

Published in: **Scientific Reports.** 2015;5: 11075.

The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis

Ralph T.H. Leijenaar, Georgi Nalbantov, Sara Carvalho, Wouter J.C. van Elmpt, Esther G.C. Troost, Ronald Boellaard, Hugo J.W.L. Aerts, Robert J. Gillies, Philippe Lambin

ABSTRACT

FDG-PET-derived textural features describing intra-tumor heterogeneity are increasingly investigated as imaging biomarkers. As part of the process of quantifying heterogeneity, image intensities (SUVs) are typically resampled into a reduced number of discrete bins. We focused on the implications of the manner in which this discretization is implemented. Two methods were evaluated: (1) R_D, dividing the SUV range into D equally spaced bins, where the intensity resolution (i.e. bin size) varies per image; and (2) R_B, maintaining a constant intensity resolution B. Clinical feasibility was assessed on 35 lung cancer patients, imaged before and in the second week of radiotherapy. Forty-four textural features were determined for different D and B for both imaging time points. Feature values depended on the intensity resolution and out of both assessed methods, R_B was shown to allow for a meaningful inter- and intra-patient comparison of feature values. Overall, patients ranked differently according to feature values–which was used as a surrogate for textural feature interpretation–between both discretization methods. Our study shows that the manner of SUV discretization has a crucial effect on the resulting textural features and the interpretation thereof, emphasizing the importance of standardized methodology in tumor texture analysis.

INTRODUCTION

In recent years, oncological research has increasingly focused on the prediction of treatment outcome based on individual patient and tumor characteristics[1], aiming to avoid the one-size-fits-all treatment approach that under- and over-treats a large number of patients. Imaging can play a crucial role here, as it allows for a non-invasive identification and characterization of the tumor [2, 3]. Positron emission tomography (PET) is a valuable tool for detecting and staging cancer [4]. In recent years, PET imaging has also been increasingly used for decision support [5], treatment planning [6, 7] and response monitoring during radiotherapy [8]. The most widely used PET tracer is [18F] fluoro-2-deoxy-D-glucose (FDG), commonly quantified by standardized uptake values (SUVs) [9]. Easily derived SUV measurements, such as the maximum or peak SUV [10], are described as predictors for treatment outcome [11-14]. Additionally, more advanced quantitative imaging features describing tumor image texture (i.e. the spatial arrangement of intensities within the image), which reflect intra-tumor heterogeneity of metabolic activity, are increasingly being investigated as potential imaging biomarkers in lung [15, 16], head and neck [17, 18], cervical [17, 19], esophageal [20-22] and other cancers [23, 24] — a field of research often referred to as 'Radiomics' [2, 3, 25-29].

Efforts have been made to provide guidelines for quality control measures in PET imaging and to standardize patient preparation, dose administration, image acquisition, image reconstruction and SUV normalization, in such a way that absolute SUV measurements are interchangeable in multicenter studies [30]. Interchangeable SUV measurements are very important in PET Radiomics, but the methodology used to determine textural features is also subject to variability. Standardization is therefore needed [31-33] (**Figure 1**).
One important methodological factor is SUV discretization (i.e. resampling image intensity values). Discretization reduces the otherwise infinite possible number of intensity values to a finite set and effectively reduces image noise. Most recent literature describes using a fixed number (e.g. 8, 16) of discrete resampled values or 'bins' to divide the image SUV range into equally spaced intervals before calculating textural features [15-17, 19-23, 32, 34-36]. Consequently, this results in discretized images with varying bin sizes or 'intensity resolutions,' depending on the SUV range. An alternative discretization method is to resample the image SUVs with a fixed bin size in units of SUV (e.g. 0.1, 0.5), maintaining a constant intensity resolution across all tumor images [37].

When aiming to identify imaging biomarkers in cohort and multicenter studies or trials, it is important that textural features and their ascribed values be directly comparable, both inter- and intra-patient, in order to derive meaningful conclusions. To our knowledge, the effect of the SUV discretization method in this respect has not been previously evaluated and we hypothesize that the aforementioned intensity resolution used for SUV discretization plays a key role in this regard.

The general objectives of our study are to compare both aforementioned conceptually different discretization methods for several popular textural features and to identify which

of these methods is most appropriate for texture quantification in a clinical setting. We will specifically investigate the role of the intensity resolution and use a clinical case study to demonstrate the effect of the SUV discretization methodology on the interpretation of the assessed textural features.

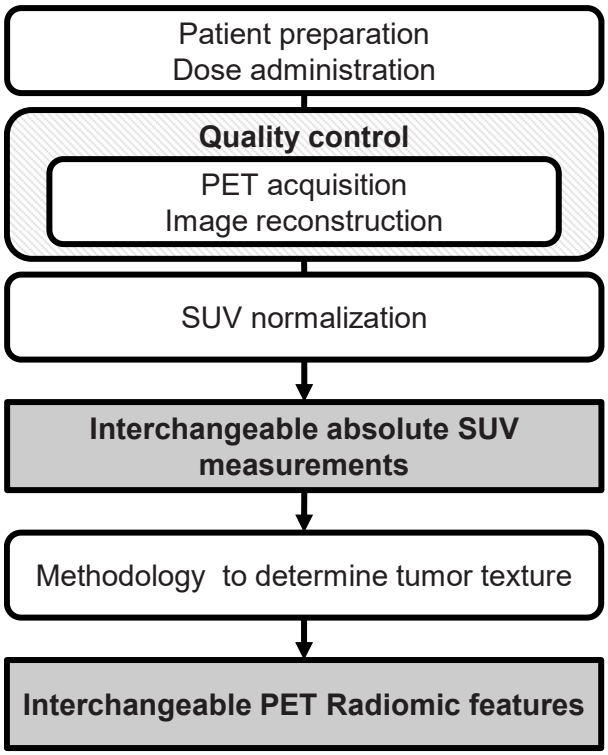


Figure 1 – Levels of standardization in PET Radiomics. Interchangeable absolute SUV measurements are obtained by standardizing patient preparation, dose administration, image acquisition, image reconstruction and SUV normalization [30]. Standardization of the methodology used for tumor texture analysis ensures interchangeable PET Radiomic features and their ascribed values.

MATERIALS AND METHODS

Patients and PET imaging

This study comprised 35 non-small cell lung cancer (NSCLC) patients who were prospectively included in a clinical trial (NCT00522639) and scheduled for radiotherapy and/or chemotherapy between July and December 2008 [11]. 18F-FDG-PET/CT imaging was performed on a Biograph 40 PET/CT scanner (Siemens Medical Solutions) twice: (1) after induction chemotherapy but before radiotherapy and (2) during the second week of radiotherapy (**Figure 2.a-b**). Patients fasted for at least six hours before imaging. The injected amount of 18F-FDG was (4×body weight [kg]+20) MBq. Patients rested 60 minutes before image acquisition. Patients' blood glucose levels were below 10 mmol/L, so no correction for blood glucose level was applied.

PET images were iteratively reconstructed using normalization- and attenuation-weighted OSEM using 4 iterations, 8 subsets and a 5 mm Gaussian filter. The resulting images had an in-plane pixel size of 4×4 mm and a 3 mm slice thickness. PET images were converted into units SUV, normalized by patient body weight [9]. Tumor volumes of interest (VOIs) were manually delineated on fused PET/CT images for treatment planning purposes. Further details are described elsewhere [11]. This study was conducted according to national laws and guidelines and approved by the appropriate local trial committee at Maastricht University Medical Center (MUMC+), Maastricht, The Netherlands. All included patients signed an informed consent form.

Image processing and feature extraction

SUVs within the VOI were first discretized using: (1) a fixed bin size (B), or intensity resolution, in units of SUV (**Figure 2.c**) and (2) a fixed number of bins (D), or discrete resampling values (**Figure 2.d**). For image I, let $I(x)$ represent the SUV of voxel x, SUV_{min} the minimum SUV in I and SUV_{max} the maximum SUV in I. Resampling SUVs into bins with an intensity resolution of B was performed using:

$$I_B(x) = \left\lceil \frac{I(x)}{B} \right\rceil - \min\left(\left\lceil \frac{I(x)}{B} \right\rceil \right) + 1 \qquad (1)$$

Where term $[\min(\lceil I(x)/B \rceil) + 1]$ ensures that the bin count starts at 1. We us the shorthand notation R_B for this resampling method. Resampling SUVs into D bins was performed using:

$$I_D(x) = \begin{cases} 1 & I(x) = SUV_{min} \\ \left\lfloor D \times \dfrac{I(x) - SUV_{min}}{SUV_{max} - SUV_{min}} \right\rfloor & \text{otherwise} \end{cases} \qquad (2)$$

Where the intensity resolution equals $(SUV_{max} - SUV_{min})/D$. This resampling method is denoted by R_D. Discretization using R_B and R_D was performed for different discretization values B (0.05, 0.1, 0.2, 0.5 and 1 [SUV]) and D (8, 16, 32, 64 and 128), respectively.

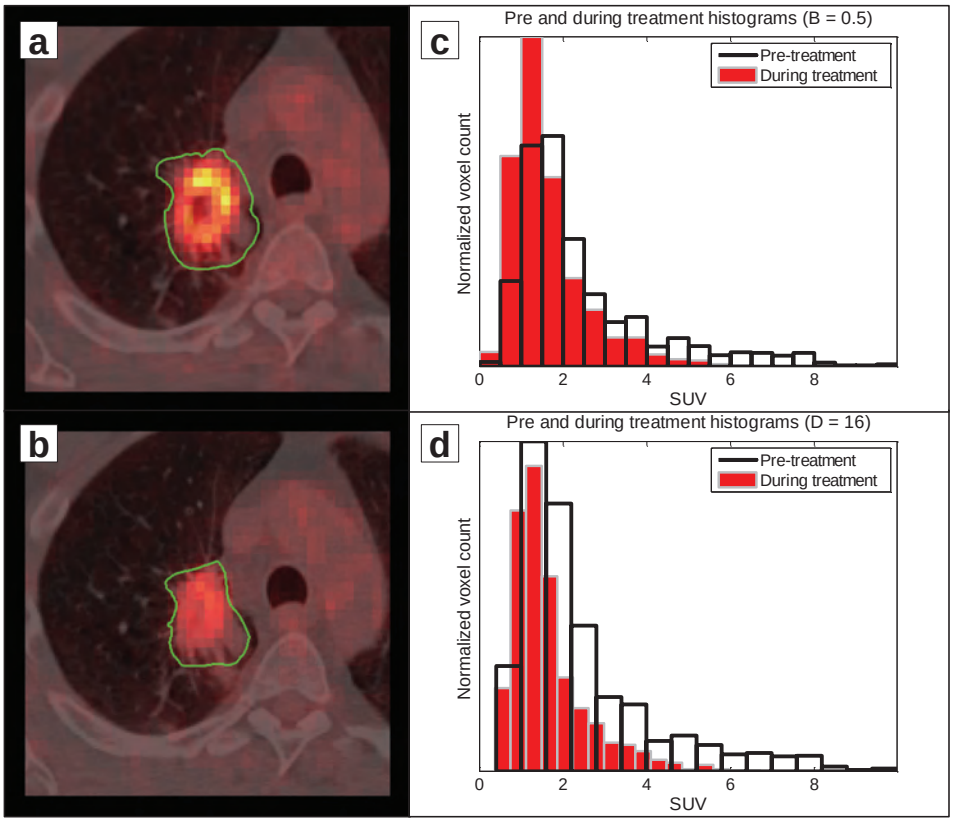


Figure 2 – ***Left column:*** Representative images of sequential imaging for one patient, showing pre-treatment imaging (**a**) and imaging during the second week of radiotherapy (**b**). The tumor delineation is outlined in green. Both images are displayed with the same window/level settings. ***Right column:*** Histograms of the pre-treatment and during treatment images, resampled with a fixed bin size (i.e. intensity resolution) (**c**) or a predefined number of bins (**d**). In (**d**), one can appreciate the difference in resulting intensity resolution when resampling with a fixed number of bins. Pre-treatment and during treatment intensity resolutions were 0.6 and 0.37 [SUV], respectively.

Textural features describing the spatial distribution of voxel intensities were calculated from gray-level co-occurrence (GLCM) [38], gray-level run-length (GLRLM) [39] and gray-level size-zone texture matrices (GLSZM) [22]. Texture matrices were determined by considering 26 connected voxels (i.e. voxels were considered to be neighbors in all 13 directions in three dimensions) at a distance of 1 voxel. Features derived from GLCM and GLRLM were calculated by averaging their value over all 13 directions. In total, 44 textural features (22 GLCM, 11 GLRLM and 11 GLSZM) were calculated. Changes in feature values between the pre-treatment and during treatment imaging time points were described as delta features, defined as:

$$\Delta X = X_{during\ treatment} - X_{pretreatment} \quad (3)$$

Image analysis was performed in Matlab R2012b (The Mathworks, Natick, MA) using an adapted version of CERR [40] and software developed in-house to extract textural features. Mathematical definitions for features assessed in this study are described elsewhere [37].

Statistical analysis

For both R_B and R_D, the pairwise intra-class correlation coefficient (ICC) [41] was calculated for each feature for all possible pairwise combinations of B (ICC_B) and D (ICC_D), to assess whether pre-treatment feature values were consistent for different discretization values. The ICC was defined as:

$$ICC = \frac{BMS - WMS}{BMS + WMS} \quad (4)$$

Where BMS and WMS are the between-subjects and within-subjects mean squares, respectively, obtained by Kruskal-Wallis one-way ANOVA. An ICC of 1 indicates perfect agreement (i.e. identical feature values).

Patient rankings according to feature value were created to serve as a surrogate for textural feature interpretation. Pairwise correlations between patient rankings were evaluated with Spearman's rank correlation coefficient (ρ). We compared patient rakings according to pre-treatment feature values and patient rankings according to delta feature values (ΔX), between all possible pairwise combinations of B (ρ^{BB}), D (ρ^{DD}) and B and D (ρ^{BD}). We considered a pairwise ρ to indicate acceptable concordance between rankings when $\rho > 0.9$. Statistical analysis was performed in Matlab R2012b.

RESULTS

Consistency of feature values for varying intensity resolutions

To assess whether feature values (using either R_B or R_D) were consistent for different discretization values, we calculated the pairwise ICCs for each feature between different values of B (ICC_B) and D (ICC_D), respectively. This analysis was performed on the pre-treatment images. For each feature, we reported the range and median of all pairwise ICCs (**Figure 3**). None of the observed pairwise ICCs was higher than 0.85, meaning that textural features and their ascribed value depend on the intensity resolution used for SUV discretization.

Variability of intensity resolution when resampling with a fixed number of bins

Using R_D, we determined the pre-treatment and during treatment bin sizes, as well as their difference, for each lesion. We observed a significant variation in both inter- and

intra-lesional intensity resolution, which is directly proportional to the SUV range. The ratio of the largest with the smallest observed intensity resolution was 1254% for pre-treatment imaging and 1038% for during treatment imaging. Absolute percentage differences in intensity resolution between pre-treatment and during treatment images ranged between 0.5% and 56%, with a median of 21%.

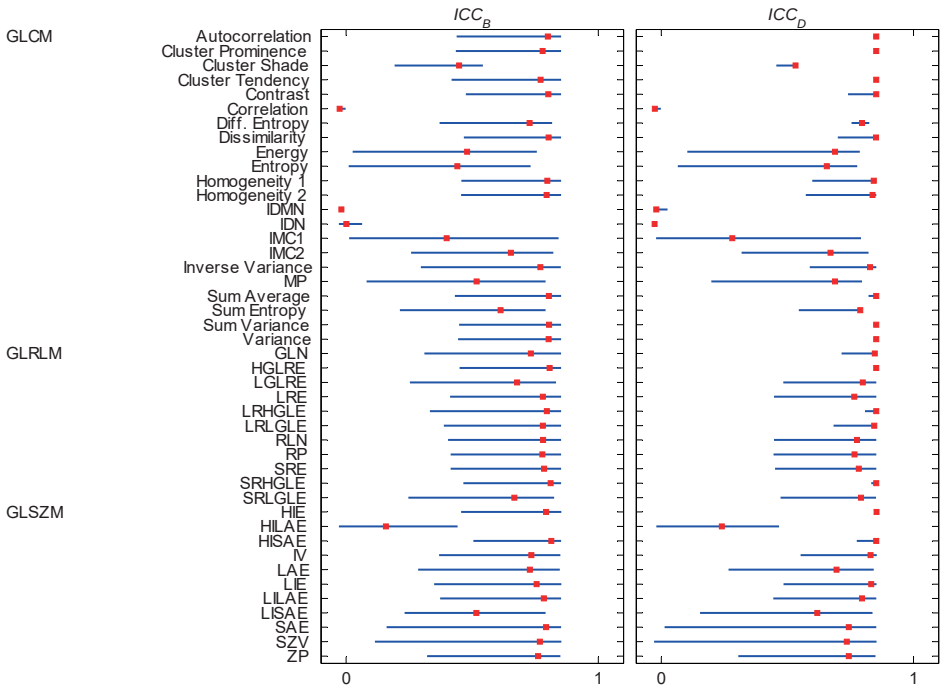


Figure 3 – Graphical representation of pairwise ICCs for each feature for different values of B (ICC_B) and D (ICC_D), based on pre-treatment imaging. Blue lines extend from the minimum to the maximum observed ICC value. Median ICC values are represented by the red markers. ***Abbreviations of feature groups:*** gray-level co-occurrence (GLCM), gray-level run-length (GLRLM) and gray-level size-zone (GLSZM). ***Abbreviations of feature names:*** Difference Entropy (Diff. Entropy), Inverse difference moment normalized (IDMN), Inverse difference normalized (IDN), Informational measure of correlation 1 (IMC1), Informational measure of correlation 2 (IMC2), Maximum probability (MP), Gray-Level Nonuniformity (GLN), High Gray-Level Run Emphasis (HGLRE), Low Gray-Level Run Emphasis (LGLRE), Long Run Emphasis (LRE), Long Run High Gray-Level Emphasis (LRHGLE), Long Run Low Gray-Level Emphasis (LRLGLE), Run-Length Nonuniformity (RLN), Run Percentage (RP), Short Run Emphasis (SRE), Short Run High Gray-Level Emphasis (SRHGLE), Short Run Low Gray-Level Emphasis (SRLGLE), High Intensity Emphasis (HIE), High Intensity Large Area Emphasis (HILAE), High Intensity Small Area Emphasis (HISAE), Intensity Variability (IV), Large Area Emphasis (LAE), Low Intensity Emphasis (LIE), Low Intensity Large Area Emphasis (LILAE), Low Intensity Small Area Emphasis (LISAE), Small Area Emphasis (SAE), Size-Zone Variability (SZV), Zone Percentage (ZP).

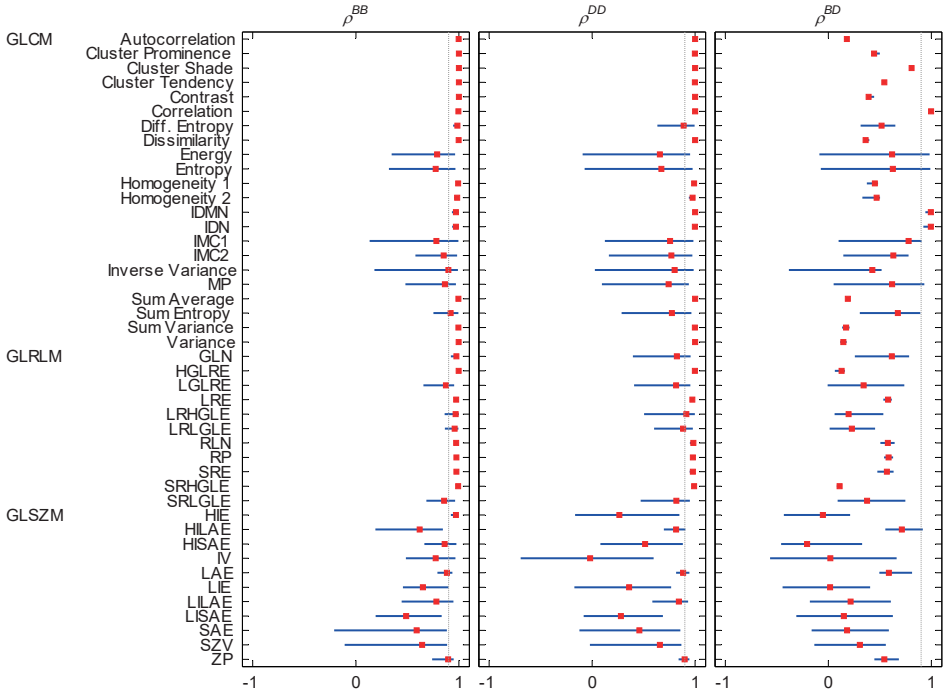


Figure 4 – Graphical representation of pairwise Spearman rank correlations between patient rankings according to feature value for different B (ρ^{BB}), different D (ρ^{DD}) and between different B and D (ρ^{BD}), based on pre-treatment imaging. Blue lines extend from the minimum to the maximum observed pairwise ρ. Median ρ values are represented by the red markers. The gray vertical line represents $\rho = 0.9$. For abbreviations, see the caption of Fig. 3.

Comparing patient rankings based on pre-treatment feature values

For each feature we determined the patient ranking according to feature value, using R_B and R_D for different resampling values B and D, respectively. We then calculated pairwise ρ of patient rankings between different B (ρ^{BB}), different D (ρ^{DD}) and between different B and D (ρ^{BD}). For each feature, we reported the range and median of all pairwise ρ (**Figure 4**). We identified 14 GLCM and 6 GLRLM features to give reliable patient rankings for both discretization methods (i.e. all pairwise $\rho^{BB} > 0.9$ and all pairwise $\rho^{DD} > 0.9$), meaning that patient rankings were nearly not affected by changes in intensity resolution. GLCM 'Difference entropy,' GLRLM 'Gray-Level Non-uniformity (GLN)' and GLSZM 'High Intensity Emphasis (HIE)' were only found to provide robust patient rankings for different resampling values when using R_B. GLCM features 'Correlation,' 'Inverse Difference Moment Normalized (IDMN)' and 'Inverse Difference Normalized (IDN)' provided very similar patient rankings between both discretization methods, regardless of the

value of either B or D (i.e. all pairwise $\rho^{BD} > 0.9$). All other features presented dissimilar patient rankings between both discretization methods.

Comparing patient rankings based on delta feature values

We also performed pairwise comparisons of patient rankings for each ΔX between different B (ρ_{Δ}^{BB}), D (ρ_{Δ}^{DD}) and different B and D (ρ_{Δ}^{BD}). For each feature, we reported the range and median of all pairwise ρ (**Figure 5**). ρ_{Δ}^{BB} and ρ_{Δ}^{DD} were both higher than 0.9 for ΔX of 10 GLCM features and 2 GLRLM features. ΔX of GLCM features 'Difference Entropy,' 'Homogeneity 1' and 'Sum Entropy' were only found to give similar patient rankings for different resampling values when using R_B. For ΔX of GLCM features 'IDMN' and 'IDN,' this was the case when using R_D for different D. The high ρ_{Δ}^{BD} (0.95–1.00) for all pairwise comparisons for GLCM feature 'Correlation' indicated highly similar patient rankings based on ΔX between both discretization methods, regardless of the value of B or D. Some pairwise ρ_{Δ}^{BD} for GLCM features 'IDMN' and 'IDN' indicated similar patient rankings for ΔX as well, but with a large range for ρ_{Δ}^{BD} (0.55–0.98 and 0.57–0.99, respectively). For all other ΔX, assessed patient rankings were found to be discordant between both discretization methods.

DISCUSSION

We compared tumor texture analysis based on SUV discretization using either a fixed number of bins (R_D) or a fixed bin size in units SUV (R_B), in the context of clinical treatment response assessment. Textural feature values were shown to depend on the intensity resolution used for SUV discretization. Overall, both resampling methods gave discordant results in terms of interpreting textural features. In the following section, we will discuss which method may be appropriate for use in a clinical setting.

Correct comparison of textural feature values

As pointed out earlier, it is important that textural feature values be directly comparable, both inter- and intra-patient, in order to derive meaningful conclusions from tumor texture analysis. The key role of the intensity resolution in this respect can be illustrated by the mathematical background of histogram bin probabilities. Let X be a continuous random variable, such as SUVs in a tumor image, with probability density function $f(x)$. The bin probabilities $P(i)$ of the first order histogram, considering equally spaced and non-overlapping bins, are then defined as:

$$P(i) = \int_{t(i)}^{t(i)+w} f(x)dx \qquad (5)$$

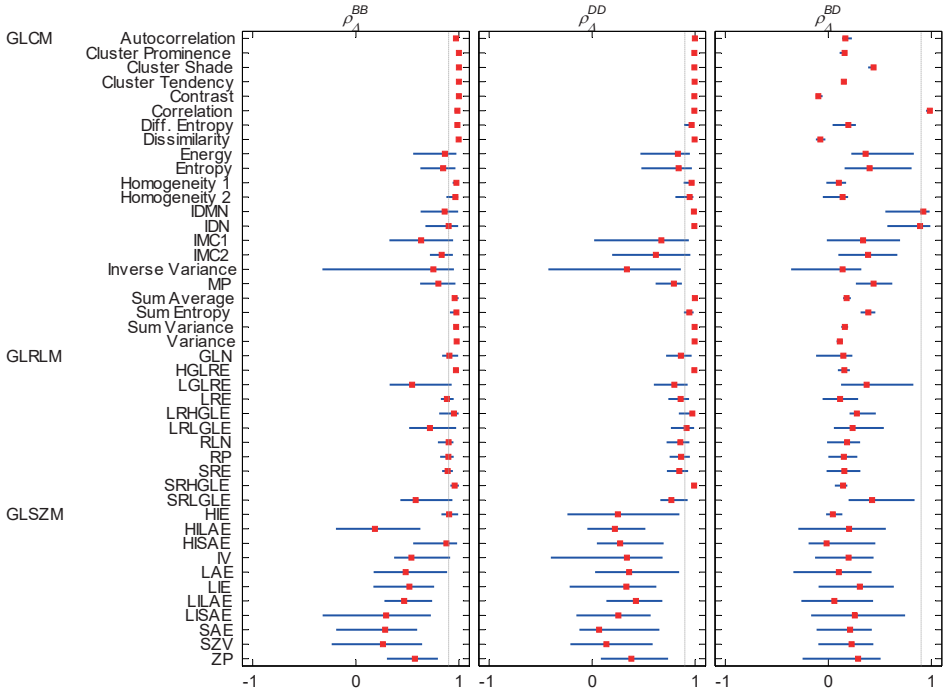


Figure 5 – Graphical representation of pairwise Spearman rank correlations between patient rankings according to ΔX for different B (ρ_Δ^{BB}), D (ρ_Δ^{DD}), and different B and D (ρ_Δ^{BD}). Blue lines extend from the minimum to the maximum observed pairwise ρ_Δ. Median ρ_Δ values are represented by the red markers. The gray vertical line represents $\rho = 0.9$. For abbreviations, see the caption of Fig. 3.

Where w represents the histogram bin size (i.e. the intensity resolution) and $t(i)$ denotes the left-hand endpoint of bin i. Analogous to $P(i)$, textural matrices are essentially histograms of joint probability densities that describe the probability of a voxel assigned to bin i being either (1) adjacent to a voxel assigned to bin j ($P_{GLCM}(i,j)$), (2) part of a consecutive run of l voxels assigned to bin i ($P_{GLRLM}(i,l)$) or (3) part of a connected neighborhood of v voxels ($P_{GLSZM}(i,v)$).

The aim is to compare textural feature values calculated from these histograms between tumor images. For all tumor images, the image intensities (x) are not dimensionless, but measured in SUV units. Maintaining a constant intensity resolution (w, in SUV units) across tumor images yields identical histogram probability definitions ($P(i)$) for each image, and hence directly comparable numerical values of each calculated feature. Using a non-constant intensity resolution across images causes one to quantify patterns (i.e. texture) on a different intensity scale (in terms of SUV) in each image.

By calculating pairwise ICCs, we observed that feature values indeed depend on the intensity resolution used for SUV discretization (**Figure 3**). More importantly, there was a significant inter- and intra-lesional variation in intensity resolution during the course of

treatment when using R_D for image intensity resampling. Effectively, R_D discards the absolute radiotracer uptake information (i.e. metabolic activity), by considering each tumor image to have the same dimensionless range of intensity. In this respect, we consider R_D to be a less appropriate choice for SUV discretization in a clinical setting, as it results in textural feature values that are not defined on the same SUV scale for each tumor image. In contrast, a constant intensity resolution is maintained across resampled images when using R_B for SUV discretization, which we believe makes it a more suitable method for tumor texture analysis.

Impact of SUV discretization method and intensity resolution on the interpretation of textural features

Several textural features were found to provide reliable patient rankings using either R_B or R_D for discretization, suggesting that results based on these features may be compared between studies if they exclusively use R_B (where studies may use a different intensity resolution B) or R_D (where studies may use a different number of bins D). However, as discussed in the previous section, we find R_D to be less appropriate in a clinical setting considering that tumor image intensities are measured in SUV units and that tumor images generally do not have the same SUV range. We therefore illustrated the implications of SUV discretization with R_D instead of R_B on the interpretation of textural features in our clinical case study. Both discretization methods resulted overall in patients being ranked differently according to their feature value (**Figure 4-5**). These results show that the manner of SUV discretization can affect the interpretation of textural features and should therefore be carefully considered in tumor texture analysis.

We furthermore observed that when R_B was used, patient rankings for several features were affected by the choice of intensity resolution (B). For those features, at least one pairwise ρ^{BB} was found to be lower than 0.9 (**Figure 4**). This suggests that results obtained for those features cannot be directly compared when different intensity resolutions are used and also suggests that their interpretation (e.g. prognostic or predictive value) depends on the intensity resolution.

It is noteworthy that the GLCM feature 'Correlation' was the only feature observed to have highly similar patient rankings over the course of treatment, regardless of the discretization method or discretization value used (**Figure 4-5**). This suggests that results obtained for this particular feature might be reliably compared between studies, provided the same discretization method is used throughout each specific study.

We used different arbitrary values for B and D throughout our study, where we kept the ratio between the smallest and largest B or D approximately the same and reasonably large. Although other values may be used as well, we found this selection to be sufficient to study our objectives. In terms of R_B however, an optimal intensity resolution cannot be straightforwardly determined. A value of 0.5 [SUV] has been described earlier, but without substantial motivation [37]. Methods for estimating an optimal intensity resolution

could be performed [42]. It should then be emphasized that the same intensity resolution needs to be maintained throughout the entire study, as determining a separate bin size for each individual patient results in non-comparable feature values. However, estimating an optimal intensity resolution does not take into account the aforementioned effect the intensity resolution has on feature interpretation, as well as the fact that using different intensity resolutions may result in complementary information [18]. In this respect, clinical validation including outcome measures is necessary to identify optimal settings that lead to meaningful results in tumor texture analysis.

Standardization in texture analysis

FDG-PET quantification is affected by several factors, including for instance breathing motion in lung [30]. Recent studies have investigated several technical aspects of FDG-PET-derived textural parameters in different cancer sites, including their test-retest repeatability and robustness regarding tumor delineation or partial volume correction [35-37], or their variability due to image acquisition and reconstruction parameters [34]. In order to provide a complete overview and acknowledging that feature stability may as well be dependent on the methodology used for SUV discretization in tumor texture analysis, we did not exclude textural features previously reported to have limited repeatability or robustness. Reliability analyses should however be performed at specific settings used in tumor texture analysis, in order to identify those features suitable for treatment assessment. The aforementioned studies point to the importance of robust and standardized PET protocols in terms of reliable quantification of tumor heterogeneity with textural features, especially when the SUV is considered to be an interchangeable quantity [30, 33]. This becomes even more essential when using fixed intensity resolutions for SUV discretization, as shown in this paper. Our study confirms that using standardized methodology for tumor texture analysis is also an important aspect of identifying and validating imaging biomarkers related to a certain outcome or underlying biology [43, 44], between different studies or trials [31, 32].

CONCLUSION

When aiming to identify and validate imaging biomarkers with tumor texture analysis of FDG-PET, it is important that the textural feature values be directly comparable, both inter- and intra-patient, in order to derive meaningful conclusions. We focused on the effect of SUV discretization and compared tumor texture analysis based on SUV discretization using a fixed intensity resolution (i.e. bin size) in units SUV (R_B) with using a fixed number of bins (R_D). We showed that maintaining a constant intensity resolution for SUV discretization across tumor images (R_B) yields textural feature values that are defined on the same SUV scale, allowing for a meaningful comparison of texture between images.

Discretizing SUVs using R_D was found to be less appropriate for inter- and intra-patient comparison of textural feature values in a clinical setting. The interpretation of textural features was overall different between both discretization methods and, for several features, affected by the choice of intensity resolution. Our study shows that the manner of SUV discretization has a crucial effect on the resulting textural features and the interpretation thereof and should therefore be carefully considered, underlining the importance of standardized methodology in tumor texture analysis.

ACKNOWLEDGMENTS

The authors acknowledge the support of the QuIC-ConCePT project, partly funded by EFPIA companies and the Innovative Medicine Initiative Joint Undertaking (IMI JU) under Grant Agreement No. 115151. Authors also acknowledge financial support from the National Institute of Health (NIH-USA U01 CA 143062-01, Radiomics of NSCLC), the CTMM framework (AIRFORCE project, grant 030-103), EU 6th and 7th framework program (EU-ROXY, METOXIA, EURECA, ARTFORCE), euroCAT (IVA Interreg - www.eurocat.info), SME Phase 2 (EU proposal 673780 – RAIL), Alpe d'HuZes-KWF (DESIGN), Kankeronderzoekfonds Limburg from the Health Foundation Limburg and the Dutch Cancer Society (KWF UM 2011-5020, KWF UM 2009-4454).

REFERENCES

[1] Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. Nat Rev Clin Oncol 2013;10: 27-40.

[2] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48: 441-446.

[3] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5: 4006.

[4] Lin P, Koh ES, Lin M, Vinod SK, Ho-Shon I, Yap J, et al. Diagnostic and staging impact of radiotherapy planning FDG-PET-CT in non-small-cell lung cancer. Radiother Oncol 2011;101: 284-290.

[5] Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CM, et al. 'Rapid Learning health care in oncology' - an approach towards decision support systems enabling customised radiotherapy'. Radiother Oncol 2013;109: 159-164.

[6] De Ruysscher D, Nestle U, Jeraj R, Macmanus M. PET scans in radiotherapy planning of lung cancer. Lung Cancer 2012;75: 141-145.

[7] Troost EG, Schinagl DA, Bussink J, Boerman OC, van der Kogel AJ, Oyen WJ, et al. Innovations in radiotherapy planning of head and neck cancers: role of PET. J Nucl Med 2010;51: 66-76.

[8] Van Elmpt W, Pottgen C, De Ruysscher D. Therapy response assessment in radiotherapy of lung cancer. Q J Nucl Med Mol Imaging 2011;55: 648-654.

[9] Thie J. Understanding the standardized uptake value, its methods, and implications for usage. J Nucl Med 2004;45: 1431-1434.

[10] Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET response criteria in solid tumors. J Nucl Med 2009;50: 122S-150S.

[11] van Elmpt W, Ollers M, Dingemans AM, Lambin P, De Ruysscher D. Response assessment using 18F-FDG PET early in the course of radiotherapy correlates with survival in advanced-stage non-small cell lung cancer. J Nucl Med 2012;53: 1514-1520.

[12] Takeda A, Yokosuka N, Ohashi T, Kunieda E, Fujii H, Aoki Y, et al. The maximum standardized uptake value (SUVmax) on FDG-PET is a strong predictor of local recurrence for localized non-small-cell lung cancer after stereotactic body radiotherapy (SBRT). Radiother Oncol 2011;101: 291-297.

[13] Velazquez ER, Aerts HJ, Oberije C, De Ruysscher D, Lambin P. Prediction of residual metabolic activity after treatment in NSCLC patients. Acta Oncol 2010;49: 1033-1039.

[14] Carvalho S, Leijenaar RT, Velazquez ER, Oberije C, Parmar C, van Elmpt W, et al. Prognostic value of metabolic metrics extracted from baseline positron emission tomography images in non-small cell lung cancer. Acta Oncol 2013;52: 1398-1404.

[15] Vaidya M, Creach KM, Frye J, Dehdashti F, Bradley JD, El Naqa I. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. Radiother Oncol 2012;102: 239-245.

[16] Cook GJ, Yip C, Siddique M, Goh V, Chicklore S, Roy A, et al. Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy? J Nucl Med 2013;54: 19-26.

[17] El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit 2009;42: 1162-1171.

[18] Cheng NM, Fang YH, Chang JT, Huang CG, Tsan DL, Ng SH, et al. Textural features of pretreatment 18F-FDG PET/CT images: prognostic significance in patients with advanced T-stage oropharyngeal squamous cell carcinoma. J Nucl Med 2013;54: 1703-1709.

[19] Yang F, Thomas MA, Dehdashti F, Grigsby PW. Temporal analysis of intratumoral metabolic heterogeneity characterized by textural features in cervical cancer. Eur J Nucl Med Mol Imaging 2013;40: 716-727.

[20] Tan S, Kligerman S, Chen W, Lu M, Kim G, Feigenberg S, et al. Spatial-Temporal [(18)F]FDG-PET Features for Predicting Pathologic Response of Esophageal Cancer to Neoadjuvant Chemoradiation Therapy. Int J Radiat Oncol Biol Phys 2012.
[21] Dong X, Xing L, Wu P, Fu Z, Wan H, Li D, et al. Three-dimensional positron emission tomography image texture analysis of esophageal squamous cell carcinoma: relationship between tumor 18F-fluorodeoxyglucose uptake heterogeneity, maximum standardized uptake value, and tumor stage. Nuclear medicine communications 2013;34: 40-46.
[22] Tixier F, Le Rest CC, Hatt M, Albarghach N, Pradier O, Metges JP, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med 2011;52: 369-378.
[23] Tixier F, Groves AM, Goh V, Hatt M, Ingrand P, Le Rest CC, et al. Correlation of Intra-Tumor 18F-FDG Uptake Heterogeneity Indices with Perfusion CT Derived Parameters in Colorectal Cancer. PLoS One 2014;9: e99567.
[24] Chicklore S, Goh V, Siddique M, Roy A, Marsden PK, Cook GJ. Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis. Eur J Nucl Med Mol Imaging 2013;40: 133-140.
[25] Cook GJR, Siddique M, Taylor BP, Yip C, Chicklore S, Goh V. Radiomics in PET: principles and applications. Clinical and Translational Imaging 2014.
[26] Naqa IE. The role of quantitative PET in predicting cancer treatment outcomes. Clinical and Translational Imaging 2014.
[27] Parmar C, Leijenaar RT, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. Scientific reports 2015;5: 11044.
[28] Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RT, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 2015;114: 345-350.
[29] Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One 2014;9: e102107.
[30] Boellaard R. Standards for PET image acquisition and quantitative data analysis. J Nucl Med 2009;50 Suppl 1: 11S-20S.
[31] Brooks FJ. On some misconceptions about tumor heterogeneity quantification. Eur J Nucl Med Mol Imaging 2013;40: 1292-1294.
[32] Orlhac F, Soussan M, Maisonobe JA, Garcia CA, Vanderlinden B, Buvat I. Tumor Texture Analysis in 18F-FDG PET: Relationships Between Texture Parameters, Histogram Indices, Standardized Uptake Values, Metabolic Volumes, and Total Lesion Glycolysis. J Nucl Med 2014;55: 414-422.
[33] Cheng NM, Fang YH, Yen TC. The promise and limits of PET texture analysis. Ann Nucl Med 2013.
[34] Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol 2010;49: 1012-1016.
[35] Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. J Nucl Med 2012;53: 693-700.
[36] Hatt M, Tixier F, Cheze Le Rest C, Pradier O, Visvikis D. Robustness of intratumour F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. Eur J Nucl Med Mol Imaging 2013.
[37] Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol 2013;52: 1391-1397.
[38] Haralick RM, Shanmugam K, Dinstein I. Textural Features of Image Classification. IEEE T Syst Man Cyb 1973;SMC-3: 610-621.
[39] Galloway M. Texture analysis using gray level run lengths. Comput Vision Graph 1975;4: 172-179.
[40] Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Med Phys 2003;30: 979-985.
[41] Shrout PE, Fleiss JL. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychol Bull 1979;86: 420-428.

[42] Izenman AJ. Recent Developments in Nonparametric Density Estimation. Journal of the American Statistical Association 1991;86: 205-224.

[43] Panth KM, Leijenaar RT, Carvalho S, Lieuwes NG, Yaromina A, Dubois L, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. Radiother Oncol 2015;116: 462-466.

[44] Hoeben BA, Starmans MH, Leijenaar RT, Dubois LJ, van der Kogel AJ, Kaanders JH, et al. Systematic analysis of 18F-FDG PET and metabolism, proliferation and hypoxia markers for classification of head and neck tumors. BMC cancer 2014;14: 130.

Chapter 5

Post-radiochemotherapy PET radiomics in head and neck cancer - the influence of radiomics implementation on the reproducibility of local control tumor models

Published in: **Radiotherapy and Oncology.** 2017.

Post-radiochemotherapy PET radiomics in head and neck cancer - the influence of radiomics implementation on the reproducibility of local control tumor models
Marta Bogowicz, Ralph T.H. Leijenaar, Stephanie Tanadini-Lang, Oliver Riesterer, Martin Pruschy, Gabriela Studer, Jan Unkelbach, Matthias Guckenberger, Ender Konukoglu, Philippe Lambin

ABSTRACT

Purpose

This study investigated an association of post-radiochemotherapy (RCT) PET radiomics with local tumor control in head and neck squamous cell carcinoma (HNSCC) and evaluated the models against two radiomics software implementations.

Materials and methods

649 features, available in two radiomics implementations and based on the same definitions, were extracted from HNSCC primary tumor region in 18F-FDG PET scans 3 months post definitive RCT (training cohort n=128, validation cohort n=50) and compared using the intraclass correlation coefficient (ICC). Local recurrence models were trained, separately for both implementations, using principal component analysis (PCA) and the least absolute shrinkage and selection operator. The reproducibility of the concordance indexes (CI) in univariable Cox regression for features preselected in PCA and the final multivariable models was investigated using respective features from the other implementation.

Results

Only 80 PET radiomic features yielded ICC>0.8 in the comparison between the implementations. The change of implementation caused high variability of CI in the univariable analysis. Both final models performed equally well in the training and validation cohorts (CI>0.7) independent of radiomics implementation.

Conclusion

The two post-RCT PET radiomic models, based on two different software implementations, were prognostic for local tumor control in HNSCC. However, 88% of the features was not reproducible between the implementations.

INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC) is one of the most common cancers worldwide with tobacco and alcohol consumption as well as HPV infection being the important risk factors. The standard of care for patients with locally advanced HNSCC is definitive radiochemotherapy (RCT). The locoregional recurrence rate is high, exceeding 50% in HPV negative oropharyngeal carcinoma and non-oropharyngeal cancers [1, 2]. A meta-analysis of post-RCT 18F-fluorodeoxyglucose positron emission tomography (18F-FDG PET) studies reported sensitivity and specificity of around 80% in respect to detection of local tumor recurrence or persistence in HNSCC [3]. Additionally, post-RCT FDG PET has been shown to correlate with overall survival [4].

Radiomics, a high throughput method for quantification of medical images, has been shown a promising input for treatment response modelling [5-9]. It is based on a comprehensive and quantitative analysis of a region of interest performed on different levels: shape, intensity, texture and filter-based analysis. Radiomics is a rapidly growing field of research. However, the studies have been predominantly performed in independent single-institution settings and consequently, the importance of workflow standardization has been indicated [5, 6].

Radiomics analysis requires several image pre-processing steps such as region of interest segmentation and extraction as well as image interpolation and discretization. These steps together with image acquisition and reconstruction parameters may influence radiomic features and therefore interchangeability of derived models (i.e. radiomic signatures) [10, 11]. Many institutions use different software packages for the analysis, which are often in-house developed. Although the implementations are based on the same mathematical definitions, it is likely that they will produce different results due to differences in implementation of algorithms as well as pre-processing [11].

To base clinical decisions on a prognostic model, its validation is required [12, 13]. Several strategies, characterized by different strength, can be used. A cross-validation is often implemented as a first step, followed by temporal validation using data from the same institution but from a different period. Finally, to achieve an unbiased validation, an external validation in an independent dataset should be performed [14]. Most of the radiomics studies have used cross-validation to quantify model performance and so far only one model has been validated in an external and independent dataset [15, 16]. Validation is usually performed by the same research group, using the same tools and methodology. However, radiomic features have been shown to vary with image acquisition parameters, pre-processing and contouring [5, 6, 10] and (to our knowledge) none of the previously published studies investigated the reproducibility of a radiomics-based prognostic model in terms of radiomics software implementation.

This study hypothesizes that the prognostic value of radiomic features is software implementation dependent. First, we investigated whether the 3 months post-RCT follow-

up 18F-FDG PET radiomics is prognostic for tumor recurrence in HNSCC. Two independent models were trained using two independent radiomics implementations and their performance was validated in a separate dataset. Subsequently, the reproducibility of these models was evaluated when their respective radiomic features were calculated with an independent software implementation.

MATERIALS AND METHODS

Imaging protocol and studied population

This retrospective analysis was approved by the local ethical commission. HNSCC patients treated with definitive radiochemotherapy were enrolled in the study (128 patients in the training and 50 patients in the validation cohort). The validation cohort consisted of patients treated in an institutional phase II prospective study (NCT01435252) with a standardized imaging protocol (the same slice thickness and reconstruction algorithm). Surgery or induction chemotherapy were exclusion criteria (biopsy allowed). The characteristic of the studied cohorts is presented in the **Table 1**. All patients underwent 18F-FDG PET/CT imaging prior to the treatment and 3 months after the end of the treatment as a standard follow-up examination. Depending on patient's body weight, an activity of 170-470 MBq of 18F-FDG was injected intravenously after the measurement of blood sugar level. The PET acquisition was preformed 60 minutes after 18F-FDG injection with a 3 minutes scanning time and 15 cm axial field-of-view at each bed position. Total acquisition time of the PET was 12-18 min. Images were reconstructed with an iterative algorithm (2D or 3D reconstruction in the training cohort and 3D reconstruction in the validation cohort) with an in-plane pixel size and the slice thickness of 2.73 – 5.47 mm and 3.27 – 4.25 mm, respectively. All data were acquired in the same center.

Image pre-processing and radiomics analysis

Tumors were semi-automatically segmented in the pre-treatment PET scans using a gradient-based method implemented in MIMVISTA (MIM Software Inc., Cleveland, OH, USA). The pre-treatment and post-treatment scans were rigidly registered and contours were transferred to post-treatment scans. To account for differences in image reconstruction grid all scans were rescaled to 5.5 mm cubic voxels using linear interpolation. This corresponds to the smallest resolution in the studied dataset.

The pre-processed images were shared between the institutions. Post-RCT metabolic heterogeneity was studied in the region of the primary tumor (**Figure 1**). Two independent software implementations were used: implementation from the University Hospital Zurich (USZ) and the MAASTRO clinic (MAASTRO). In total 649 features, which were based on the same definition and available in both implementations, were extracted:

- Shape (n = 8)
- Intensity-based (n = 17)
- Texture: the Gray Level Co-occurrence Matrix (GLCM; n = 24), the Neighborhood Gray Tone Difference Matrix (NGLTDM; n = 4), the Gray-Level Size Zone Matrix (GLSZM; n = 14), the Gray-Level Run Length Matrix (GLRLM; n = 14).
- Filter-based: Wavelet coiflet (n = 568).

A bin size of 0.5 SUV was used for image intensity discretization. The consistency of radiomic features calculated in two different implementations was studied using the two-way mixed single measures intraclass correlation coefficient (ICC).

Table 1 – Detailed characteristic of studied cohorts.

		Training cohort	Validation cohort
	Total number of patients	128	50
	Median follow-up (months)	46 (3-156)	16 (3-28)
	Number of local recurrences	38 (30%)	13 (26%)
Tumor stage	T1/T2	43 (34%)	6 (12%)
	T3/T4	85 (66%)	44 (88%)
HPV status	Positive	31 (24%)	22 (44%)
	Negative	36 (28%)	28 (66%)
	Unknown	61 (48%)	0
Tumor site	Oropharynx	91 (71%)	29 (58%)
	Hypopharynx	22 (17%)	7 (14%)
	Larynx	11 (9%)	7 (14%)
	Oral cavity	4 (3%)	7 (14%)
Treatment	Radiotherapy	on average 70 Gy (68 – 72 Gy)	70 Gy
	Chemotherapy	Cisplatin (40 mg/m2, up to 7 cycles) or cetuximab (loading dose 400 mg/m2 followed by 250 mg/m2 weekly)	cisplatin/cetuximab (weekly same doses as in training cohort) with or without consolidation cetuximab (500 mg/m2 biweekly x 6)
PET scanners	GE Discovery STE	64 (50%)	39 (78%)
	GE Discovery 690	10 (8%)	6 (12%)
	GE Discovery RX	23 (18%)	5 (10%)
	GE Discovery HR	15 (12%)	
	GE Discovery LS	16 (12%)	

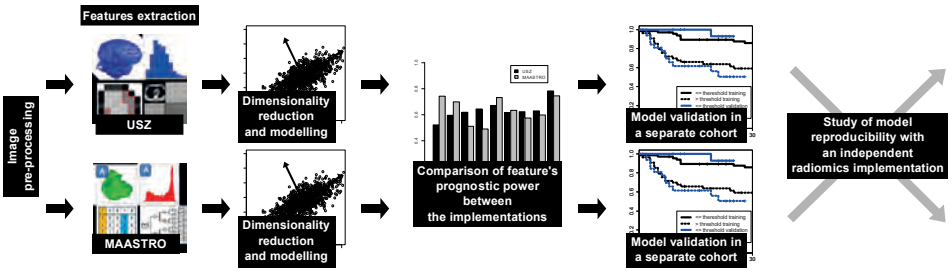


Figure 1 – Scheme of the reproducibility analysis of the local tumor control models using two independent radiomics implementation.

Features preselection and comparison of the features' prognostic power between the radiomics implementations

The following feature selection procedure was used. First, a principal component analysis (PCA) was performed to account for inter-feature correlations. The number of retained components was adjusted to represent 95% of data variance. Next, for each principal component one feature was selected to represent it. To that end we determined the feature that correlated the most (the largest Pearson correlation coefficient) with the principal component.

The prognostic power of radiomic features selected in different implementations was investigated in a univariable Cox regression. The models were fitted separately for the USZ and MAASTRO implementations. To quantify the discriminative power of different models the concordance indexes (CI) were calculated and compared between the implementations. The p-value from Cox regression was corrected for the multiple testing using false discovery rate (FDR) < 10% and the number of features defined in the PCA. The statistical analysis was performed in R (v. 3.2.3).

Prediction of local tumor recurrence and model reproducibility between the implementations

To train a final model for the association of the radiomic features derived from post-RCT PET with the likelihood of tumor recurrence, the least absolute shrinkage and selection operator (LASSO) (100 times 5-fold cross-validated) was used for variable selection in multivariable Cox regression. A random sampling with replacement was used to create a different training set in each of the LASSO iterations. In the final model we included only radiomic features with selection rate higher than 70% among all random training sets. Only the features preselected in the PCA were used in the multivariable analysis. Patients were stratified into low- and high-risk of recurrence groups based on a threshold from

the receiver operating characteristic curve for local recurrence at 18 months. The threshold was selected to equate the level of sensitivity and specificity. The groups were compared using G-rho test (p-value < 0.05). Two models were trained separately, one on the USZ and one on the MAASTRO feature set. Both models were validated in the independent cohort of patients.

Each trained model, based on the features calculated in one implementation, was later evaluated by calculating its respective features with the other independent implementation (Figure 1). The regression coefficients of the Cox model and the stratification threshold were then fixed. Model performance was quantified using the concordance index (CI). Additionally, the calibration of the models was investigated by calculation of the calibration slope based on the prognostic index [17]. The calibration slope equals 1 evidences the same level of discrimination in the training and validation datasets. Finally, the correlation of hazards obtained with two implementations and the reproducibility of the patients risk group assignments were investigated.

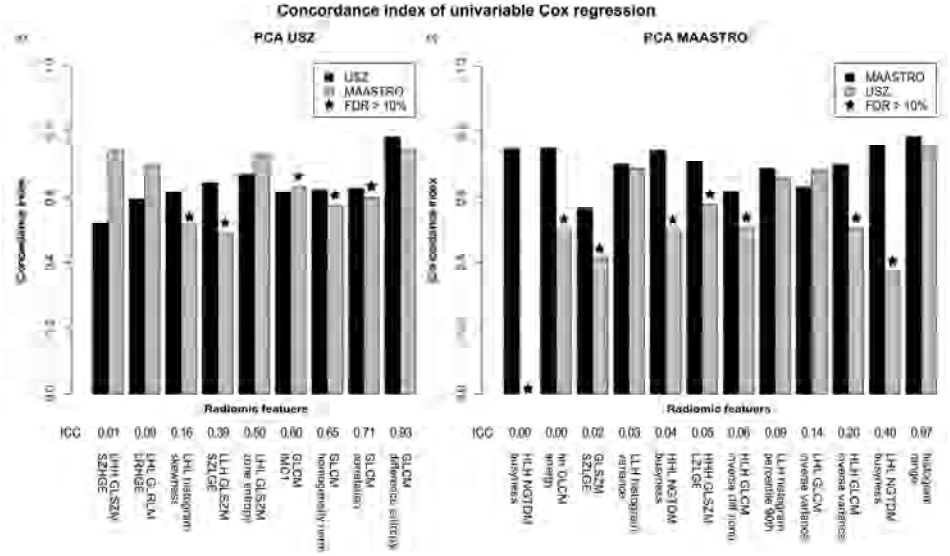


Figure 2 – Comparison of the features' prognostic power in the univariable Cox regression between the radiomics implementations. The fit was considered non-significant if false discovery rate (FDR) > 10%. The concordance indexes for the same feature varied between the radiomics implementations and this effect did not depend on the feature's intraclass correlation (ICC) from the implementations comparison. The LLL, LLH, LHL, LHH, HHH, HHL, HLH, HLL – denote the combination of wavelet filters in 3D (L – low-pass, H – high-pass).

RESULTS

Radiomic features reproducibility between the two implementations

The intraclass correlation coefficient was used to investigate features reproducibility. Out of 649 features, 46 and 80 were characterized by an ICC greater than 0.9 and 0.8, respectively. These were mostly histogram-based (92% of the features in the studied group based on the ICC > 0.8) and texture-based (68%) features calculated on the non-transformed images. The shape features showed intermediate reproducibility (50%), whereas the biggest discrepancy was observed for the wavelet features (supplement **Figure 1S**). The wavelet features where high-pass filter was applied more than once were the least reproducible. A translation of the 0.5 SUV bin size to the wavelet coefficients was different between the implementations. It resulted in a different number of analyzed gray levels in the wavelet maps (supplement **Figure 3S**). Additionally, the MAASTRO implementation uses undecimated transform, whereas the USZ implementation uses the decimated one. It influenced the resolution of analyzed maps.

Comparison of the features' prognostic power between the radiomics implementations

In the principal component analysis, 31 and 33 components retained the 95% of data variance in the USZ and MAASTRO implementation, respectively. We found only 6 representative features based on the principal components analysis to be the same for both implementations. In a univariable Cox regression, 9 features in USZ and 12 features in MAASTRO implementation yielded a FDR < 10%. Among those features, more than 50% was not significant in the univariable Cox regression when calculated with the other independent implementation (**Figure 2**). Even if the feature was significant in both implementations, a substantial difference in CI was observed. The reproducibility of a single feature's prognostic power did not depend on the value of ICC (Wilcoxon test p-value < 0.05), except of the features with almost perfect agreement.

Prediction of local tumor recurrence and comparison between the implementations

In the multivariable analysis, GLCM difference entropy was found to be prognostic in the USZ implementation, whereas the histogram range was selected from the MAASTRO implementation. Radiomic features in the final local tumor recurrence models showed high level of reproducibility between the radiomics implementations (ICC > 0.9). A strong correlation (r > 0.9) between GLCM difference entropy and histogram range was observed independent of the implementation.

Both models showed similar prognostic power in the training (5-fold cross-validation) and validation cohorts with CI ranging between 0.70 and 0.76 (**Table 2**) and allowed for a significant risk group stratification (**Figure 3**). In the validation cohort, the calibration slope was not significantly different from 1, indicating the preservation of model discriminative power (**Table 2**). Additionally, the models were prognostic in the group of HPV negative patients (supplement **Figure 4S**). In both models, tumors with higher risk of recurrence were characterized by a higher post-treatment metabolic heterogeneity (supplement **Figure 5S**).

The main research question asked in this work was to investigate the model performance when an independent radiomics implementation was used to calculate the hazards. Also in this case, the studied PET radiomics models achieved a very similar performance in terms of the concordance index as well as similar calibration slope (**Table 2**). It showed that the general discriminative power of the models was not affected by the change of the implementation. On the patient level, a strong correlation was observed between patient rankings based on the features from both implementations ($r > 0.9$). Most of the patients (around 90%) were correctly classified into low- or high-risk of recurrence group when the independent implementation was used (**Figure 4**). However, it is important to note that the reproducibility of the models is a consequence of the high ICC between the implementations for the selected features.

Table 2 – Performance of PET radiomics models for prediction of local tumor control and the stability of radiomic features between two radiomics implementations (USZ and MAASTRO). Underlined values indicates results where the same implementation was used for the training of the model and model performance evaluation, * 95% confidence interval.

		Model developed using radiomic features from	
		MAASTRO	USZ
Radiomic features		Histogram range	GLCM difference entropy
Intraclass correlation		0.97	0.93
Concordance Index			
MAASTRO features	training	0.76	0.75
	validation	0.73	0.73
USZ features	training	0.75	0.74
	validation	0.71	0.72
Calibration slope			
MAASTRO features		1.20 (0.39 – 2.02)*	1.04 (0.27 – 1.95)*
USZ features		1.13 (0.39 – 1.88)*	1.02 (0.20 - 1.83)*

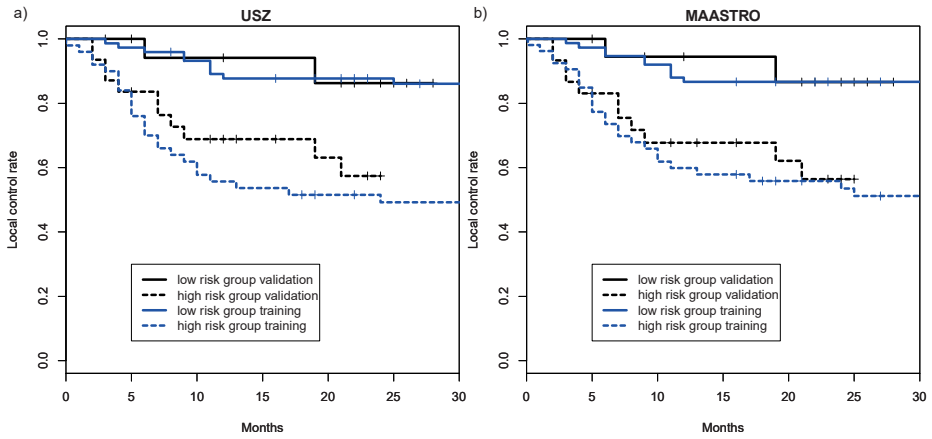


Figure 3 – PET radiomics-based local tumor recurrence models: a) USZ implementation, b) MAASTRO implementation. Local control rate curves split significantly (G-rho test p-value < 0.05) in both training and validation cohorts based on the optimal sensitivity-specificity thresholds at 18 months.

DISCUSSION

This study investigated the prognostic value of post-RCT PET radiomics in head and neck squamous cell carcinoma and tested the reproducibility of the models between independent radiomics implementations (USZ or MAASTRO). Independent of the radiomics implementation used for model training, the prognostic model for local tumor control showed a good discriminative power with a concordance index higher than 0.7 in both training and validation cohorts. Both models significantly stratified patients into low- and high-risk of recurrence groups. Furthermore, the validation of the models using an independent radiomics implementation resulted in a similar concordance indexes. However, in the modelling process we have observed that the discriminative power of single radiomic features preselected for the multivariable analysis depended on the radiomics implementation.

The value of post-treatment FDG-PET imaging for assessment of residual disease is currently unclear [3]. Recently, it has been shown in a prospective study that the positive findings on 3 months post-treatment FDG PET are a prognostic factor for overall survival and cancer-specific survival [4]. Additionally, our work shows that the heterogeneity of 3 months post-RCT FDG activity in the region of primary tumor is related to the risk of tumor recurrence. Higher histogram range (range of SUV in the region of primary tumor) and higher GLCM difference entropy corresponded to higher risk of tumor recurrence. We have further shown that these radiomics models can also significantly stratify the HPV negative patients, who belong to a group with a generally bad prognosis. The intensity-volume histogram features in the pre-treatment FDG-PET were previously shown to give

even a better prediction of tumor control in head and neck cancer (AUC = 1), however these results were obtained on a small cohort of patients with no validation (n = 9) [18].

Our prognostic models were trained on a heterogeneous dataset, different PET scanners and reconstruction algorithms were used. However, we were able to validate obtained results on a dataset with a standardized imaging protocol (the same slice thickness and reconstruction algorithm). Our findings should be further validated in datasets from other centers as the lack of calibration between different PET scanners can affect the performance of the models [19]. Additionally, we have defined our region on interest based on the pre-treatment PET images and propagated it to the post-treatment scan. The model reproducibility should be tested against different registration methods for propagation of the delineated tumor volume.

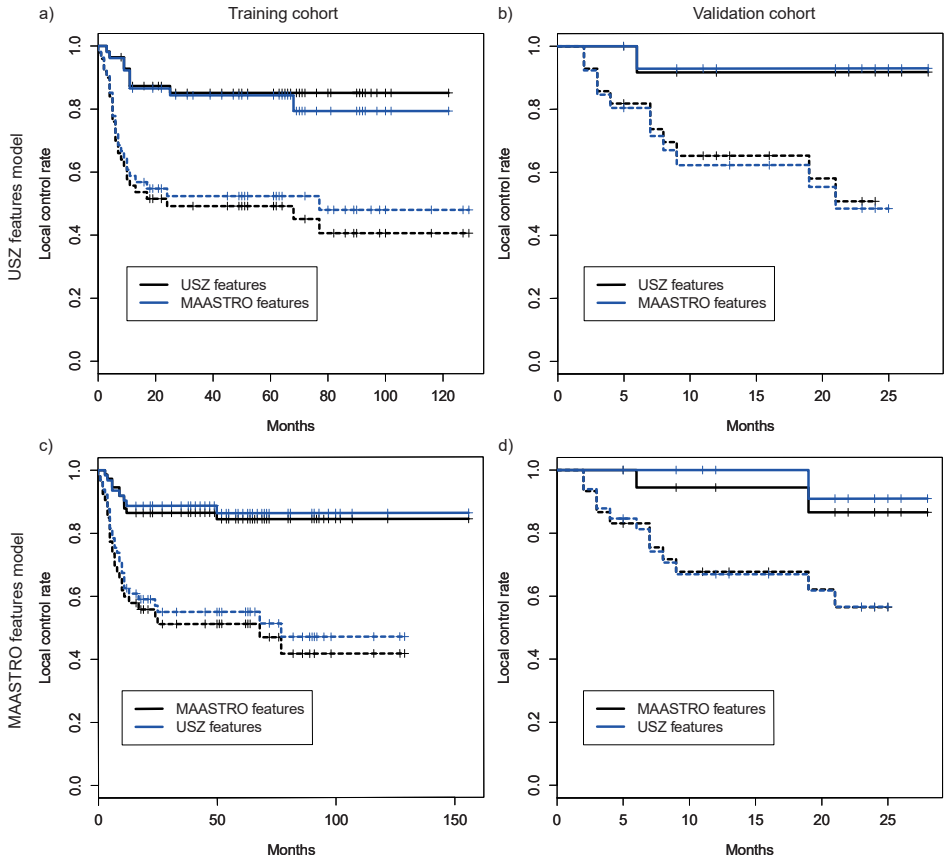


Figure 4 – Local control rate curves for low- and high-risk of recurrence groups based on the two PET radiomic models. The curves split significantly independent of the implementation (G-rho test p-value < 0.05).

The two radiomics implementations used in this study are based on the same mathematical definition of radiomic features. Additionally, the image pre-processing (image and region of interest resizing) was performed independent of the radiomics implementation and the same bin size was used for image discretization [20]. Nevertheless, a relevant variability in radiomic features value was observed, mostly for the shape and wavelet features. It was most probably caused by differences in mask extraction and wavelet transform workflow: especially in the translation of the bin size to the wavelet transformed images and the type of transform (decimated vs undecimated). Variations of contour masks extracted from the same DICOM files is also a well-known issue in different treatment planning systems [21]. The comparison of the number of analyzed voxels, as well as minimum, maximum and mean value in the GTV between two radiomics implementation is shown in **Figure 2S**. The GTV extracted with USZ implementation was always larger then in MAASTRO implementation and consequently the minimum SUV in USZ implementation was always lower. Regarding wavelet transform, the two implementations transferred differently the bin size of 0.5 SUV into the wavelet coefficients space, which resulted in different number of analyzed gray levels (**Figure 3S**). A separate study investigating a discriminative power of wavelet features obtained with the two gray level discretization methods should be conducted to justify which method is more suitable for medical image analysis. This study points out differences in radiomics workflow steps, which are often not even described in radiomics studies. Therefore, clear guidelines, such as Image Biomarker Standardization Initiative [22], providing detailed description of radiomics workflow and implementation are needed. Our final prognostic models for local tumor recurrence were reproducible when the features from the independent radiomics implementation were used, which can be explained by the fact that both models consisted of radiomic features with a high ICC in the comparison between the implementations (ICC > 0.9). Most of the available radiomic features showed a much lower agreement in this comparison. A large variation in concordance indexes was observed for the radiomic features preselected in the principal component analysis. Most of the features preselected in one implementation was not significantly associated with local tumor recurrence in the other implementation in the univariable analysis. This shows, for the first time on a clinically relevant model and dataset, that a model developed by one institution cannot be directly transferred to another center, which uses a different radiomics implementation. We recommend that each model, additionally to a detailed description of the radiomics implementation, should be published with a sample dataset and corresponding radiomics signature, such as a recently published digital phantom [9]. This will allow for a comparison of results obtained from a model, before it will be used in a prospective cohort.

In conclusion, this study shows the potential of post-RCT FDG-PET radiomics for early identification of patients with a high risk of local tumor recurrence. It also raises an awareness of the impact of radiomics software implementation on model reproducibility.

ACKNOWLEDGEMENTS

The project was supported by the Clinical Research Priority Program Tumor Oxygenation of the University of Zurich, by a grant of the Matching Fund of the University of Zurich. The clinical study used as validation dataset was supported by a research grant from Merck (Schweiz) AG. Additionally, authors acknowledge financial support from the ERC advanced grant (ERC-ADG-2015, n° 694812 - Hypoximmuno), the QuIC-ConCePT project (IMI JU; grant no. 115151). This research is also supported by the Dutch technology Foundation STW (grant n° 10696 DuCAT & n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs. Authors also acknowledge financial support from the EU 7th framework program (ARTFORCE - n° 257144, REQUITE - n° 601826), SME Phase 2 (EU proposal 673780 – RAIL), EUROSTARS (DART), the European Program H2020-2015-17 (BD2Decide - PHC30-689715 and ImmunoSABR - n° 733008), Interreg V-A Euregio Meuse-Rhine ("Euradiomics"), Alpe d'HuZes-KWF (DESIGN), Kankeronderzoekfonds Limburg from the Health Foundation Limburg, the Zuyderland-MAASTRO grant and the Dutch Cancer Society.

REFERENCES

[1] Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tan PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. The New England journal of medicine 2010;363: 24-35.

[2] Lassen P, Primdahl H, Johansen J, Kristensen CA, Andersen E, Andersen LJ, et al. Impact of HPV-associated p16-expression on radiotherapy outcome in advanced oropharynx and non-oropharynx cancer. Radiother Oncol 2014;113: 310-316.

[3] Gupta T, Master Z, Kannan S, Agarwal JP, Ghsoh-Laskar S, Rangarajan V, et al. Diagnostic performance of post-treatment FDG PET or FDG PET/CT imaging in head and neck cancer: a systematic review and meta-analysis. Eur J Nucl Med Mol Imaging 2011;38: 2083-2095.

[4] Kim SA, Roh JL, Kim JS, Lee JH, Lee SH, Choi SH, et al. 18F-FDG PET/CT surveillance for the detection of recurrence in patients with head and neck cancer. Eur J Cancer 2017;72: 62-70.

[5] Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present... any future? Eur J Nucl Med Mol Imaging 2017;44: 151-165.

[6] Yip SS, Aerts HJ. Applications and limitations of radiomics. Phys Med Biol 2016;61: R150-166.

[7] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48: 441-446.

[8] Riesterer O, Nesteruk M, Studer G, Guckenberger M, Lang S. Predictive Value of Radiomics Analysis for Local Tumor Control After Radiochemotherapy in Patients With Head and Neck cancer. International Journal of Radiation Oncology • Biology • Physics;96: S117.

[9] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol In press.

[10] Larue RT, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. Br J Radiol 2017;90: 20160665.

[11] Court LE, Fave X, Mackin D, Lee J, Yang J, Zhang L. Computational resources for radiomics. Translational Cancer Research 2016;5: 340-348.

[12] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. British journal of cancer 2015;112: 251-259.

[13] Rios Velazquez E, Hoebers F, Aerts HJ, Rietbergen MM, Brakenhoff RH, Leemans RC, et al. Externally validated HPV-based prognostic nomogram for oropharyngeal carcinoma patients yields more accurate predictions than TNM staging. Radiother Oncol 2014;113: 324-330.

[14] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ (Clinical research ed.) 2009;338.

[15] Leijenaar RT, Carvalho S, Hoebers FJ, Aerts HJ, van Elmpt WJ, Huang SH, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta Oncol 2015;54: 1423-1429.

[16] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5: 4006.

[17] Royston P, Altman D. External validation of a Cox prognostic model: principles and methods. BMC Medical Research Methodology 2013;13: 33.

[18] El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit 2009;42: 1162-1171.

[19] Adams MC, Turkington TG, Wilson JM, Wong TZ. A Systematic Review of the Factors Affecting Accuracy of SUV Measurements. American Journal of Roentgenology 2010;195: 310-320.

[20] Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Scientific reports 2015;5: 11075.

[21] Kulkarni BS, Dutt Sharma S, Hansen V, Sresty N, M S, Kandan M, et al. A prospective study of OAR volume variations between two different treatment planning systems in radiotherapy. ed.; 2015.
[22] Zwanenburg A. EP-1677: Multicentre initiative for standardisation of image biomarkers. Radiotherapy and Oncology 2017;123: S914-S915.

SUPPLEMENTARY FIGURES

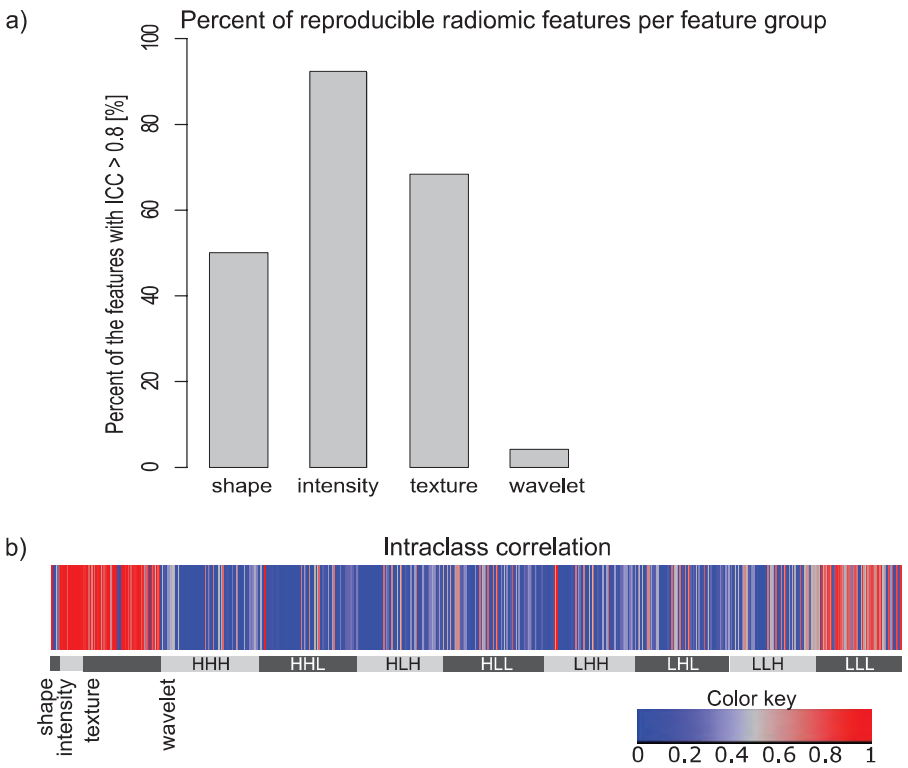


Figure 1S – The reproducibly of radiomics features between the implementations. a) Histogram of reproducible radiomic features. The intensity and texture features showed a high level of agreement between the implementations. The irreproducibility of shape and wavelet features was caused by differences in mask extraction and wavelet maps normalization. b) Intraclass correlation coefficient for different features, H – high-pass filet, L – low-pass filter. The features where high-pass filter was applied more than once were the least reproducible. The difference between the implementation in wavelet maps normalization did not have a big influence on the low-pass filtered images.

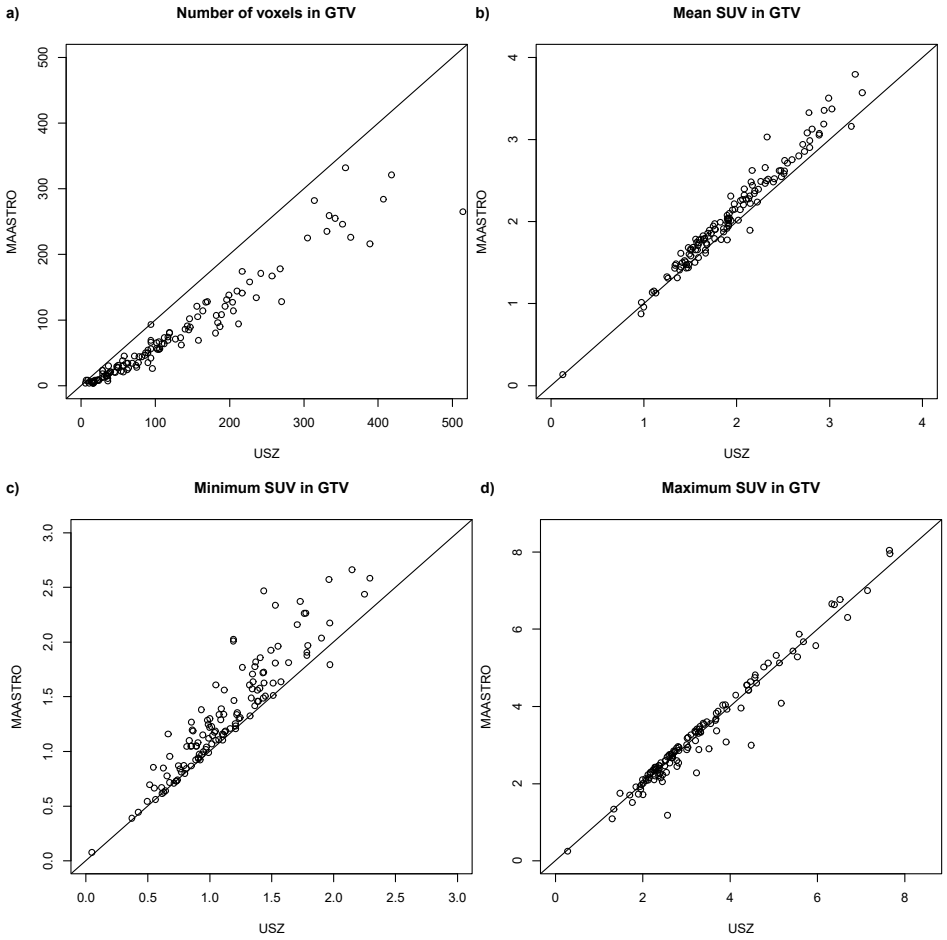


Figure 2S – The comparison of mask extraction algorithms used by the different radiomics implementation. Each point corresponds to one patient in the training cohort. The masks extracted with USZ implementation were generally larger (larger number of analyzed voxels (a)). It also resulted in a lower minimum SUV (c) observed in USZ implementation, whereas the maximum (d) and mean SUV (b) were less affected. A 1:1 line was plotted for the comparison.

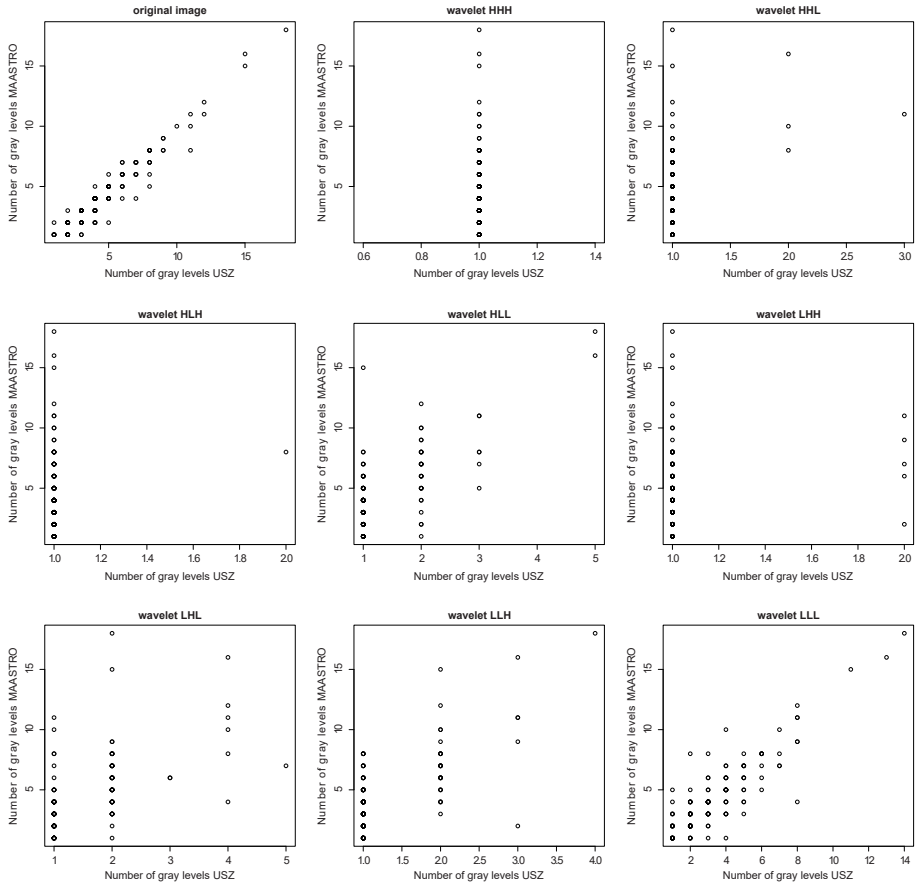


Figure 3S – The comparison of the number of gray level used in the analysis by the two independent implementations (USZ and MAASTRO). Each point corresponds to one patient in the training cohort. Although, both implementations used the same bin size (0.5 SUV) its translation to the bin size in the wavelet transformed maps differed, resulting in different number of gray levels.

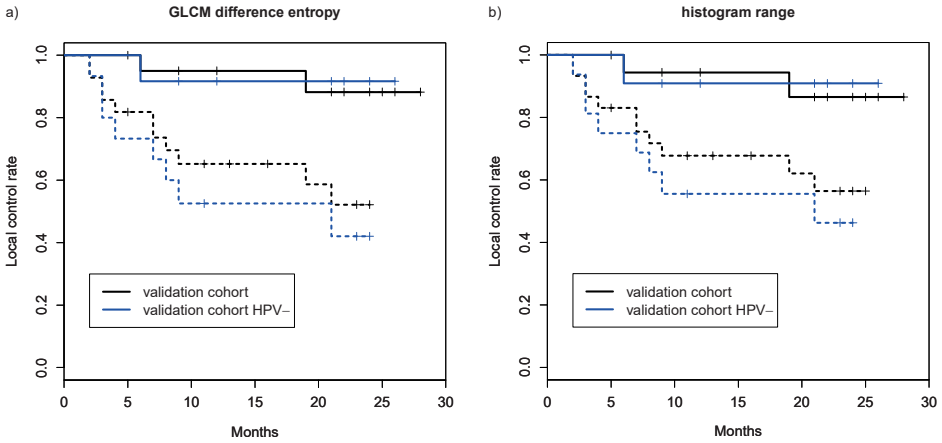


Figure 4S – PET radiomics-based local tumor recurrence models: a) USZ implementation, b) MAASTRO implementation. The models are prognostic not only in the validation cohort but also in the subgroup of HPV negative patients (CI_{USZ} = 0.78, $CI_{MAASTRO}$ = 0.82). The local control curves split significantly (G-rho test p-value < 0.05) based on the optimal sensitivity-specificity thresholds at 18 months.

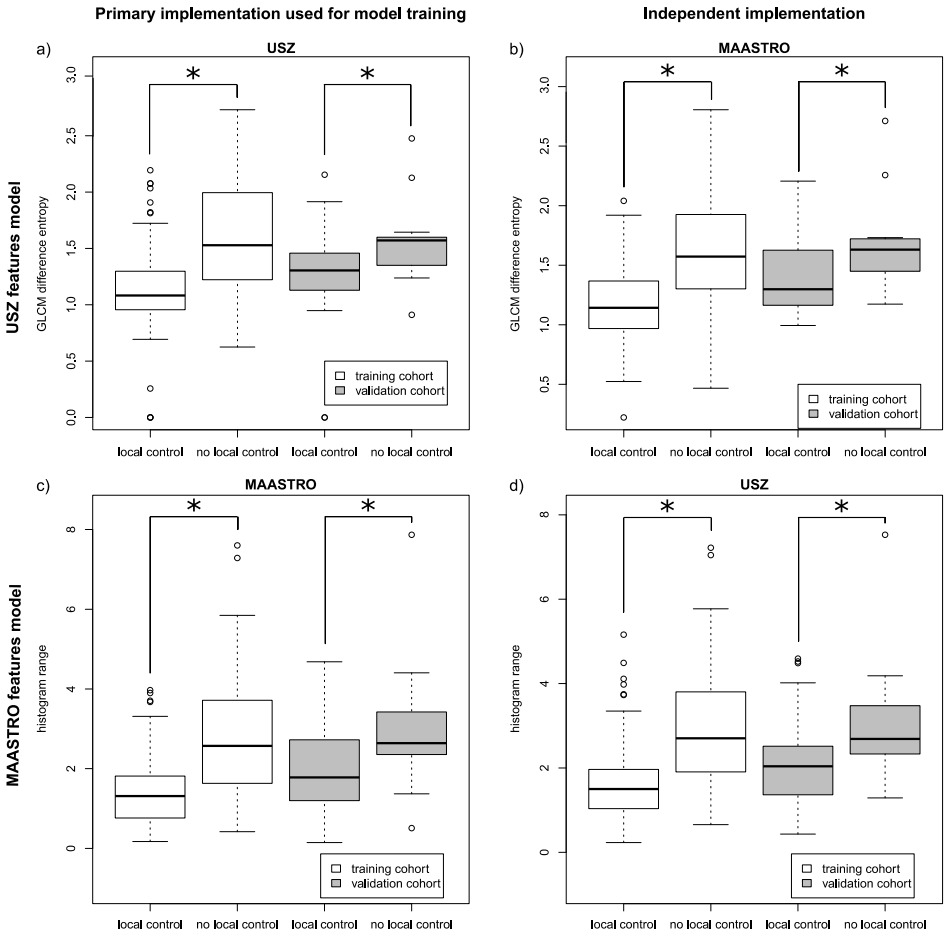


Figure 5S – Post-RCT PET radiomics signature prognostic for local tumor recurrence. In both implementations radiomic features selected in the final models were significantly different for patients with controlled tumors and with recurrences (* Wilcoxon rank-sum test $p < 0.05$).

Chapter 6

Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach

Published in: **Nature Communications**. 2014;5: 4006.

Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach

Hugo J.W.L. Aerts*, Emmanuel Rios Velazquez*, Ralph T.H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Cavalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Joseph O. Deasy, Andre Dekker, John Quackenbush, Robert J. Gillies, Philippe Lambin

*These authors contributed equally to this work

ABSTRACT

Human cancers exhibit strong phenotypic differences that can be visualized noninvasively by medical imaging. Radiomics refers to the comprehensive quantification of tumour phenotypes by applying a large number of quantitative image features. Here we present a radiomic analysis of 440 features quantifying tumour image intensity, shape and texture, which are extracted from computed tomography data of 1,019 patients with lung or head-and-neck cancer. We find that a large number of radiomic features have prognostic power in independent data sets of lung and head-and-neck cancer patients, many of which were not identified as significant before. Radiogenomics analysis reveals that a prognostic radiomic signature, capturing intratumour heterogeneity, is associated with underlying gene-expression patterns. These data suggest that radiomics identifies a general prognostic phenotype existing in both lung and head-and-neck cancer. This may have a clinical impact as imaging is routinely used in clinical practice, providing an unprecedented opportunity to improve decision-support in cancer treatment at low cost.

INTRODUCTION

Medical imaging is one of the major factors that have informed medical science and treatment. By assessing the characteristics of human tissue non-invasively, imaging is often used in clinical practice for oncologic diagnosis and treatment guidance [1-3]. A key goal of imaging is 'personalized medicine', were treatment is increasingly tailored based on specific characteristics of the patient and their disease [4].

Much of the discussion of personalized medicine has focused on molecular characterization using genomic and proteomic technologies. However, as tumours are spatially and temporally heterogeneous, these techniques are limited. They require biopsies or invasive surgeries to extract and analyse what are generally small portions of tumour tissue, which do not allow for a complete characterization of the tumour. Imaging has great potential to guide therapy because it can provide a more comprehensive view of the entire tumour and it can be used on an on-going basis to monitor the development and progression of the disease or its response to therapy. Further, imaging is non-invasive and is already often repeated during treatment in routine practice, on the contrary of genomics or proteomics, which are still challenging to implement into clinical routine.

The most widely used imaging modality in oncology is x-ray computed tomography (CT), which assesses tissue density. Indeed, CT images of lung cancer tumours exhibit strong contrast reflecting differences in the intensity of a tumour on the image, intra tumour texture, and tumour shape (**Figure 1a**). However, in clinical practice, tumour response to therapy is only measured using 1 or 2 dimensional descriptors of tumour size (RECIST and WHO, respectively) [5]. Although a change in tumour size can indicate response to therapy, it often does not predict overall or progression free survival [6, 7]. Although some investigations have characterized the appearance of a tumour on CT images, these characteristics are typically described subjectively and qualitative ("moderate heterogeneity", "highly spiculated", "large necrotic core"). However, recent advances in image acquisition, standardization, and image analysis, allow for objective and precise quantitative imaging descriptors that could potentially be used as non-invasive prognostic or predictive biomarkers.

Radiomics is an emerging field that converts imaging data into a high dimensional mineable feature space using a large number of automatically extracted data-characterization algorithms [8, 9]. We hypothesize that these imaging features capture distinct phenotypic differences of tumours and may have prognostic power and thus clinical significance across different diseases. Here we assess the clinical relevance of 440 radiomic features, many of which currently have no known clinical significance, in seven independent cohorts consisting of 1,019 lung cancer and head-and-neck cancer patients. Two data sets are used to assess the stability of the features, four data sets to assess the prognostic value of radiomic features on lung cancer patients and head-and-neck cancer patients, and one data set for association with gene-expression profiles of lung cancer patients (**Figure 2**). Our results reveal that radiomics data contain strong prognostic information in both lung

and head-and-neck cancer patients, and are associated with the underlying gene-expression patterns. These results suggest that radiomics decodes a general prognostic phenotype existing in multiple cancer types. Radiomics can have a large clinical impact, as imaging is used in routine practice worldwide, providing a method that can quantify and monitor phenotypic changes during treatment.

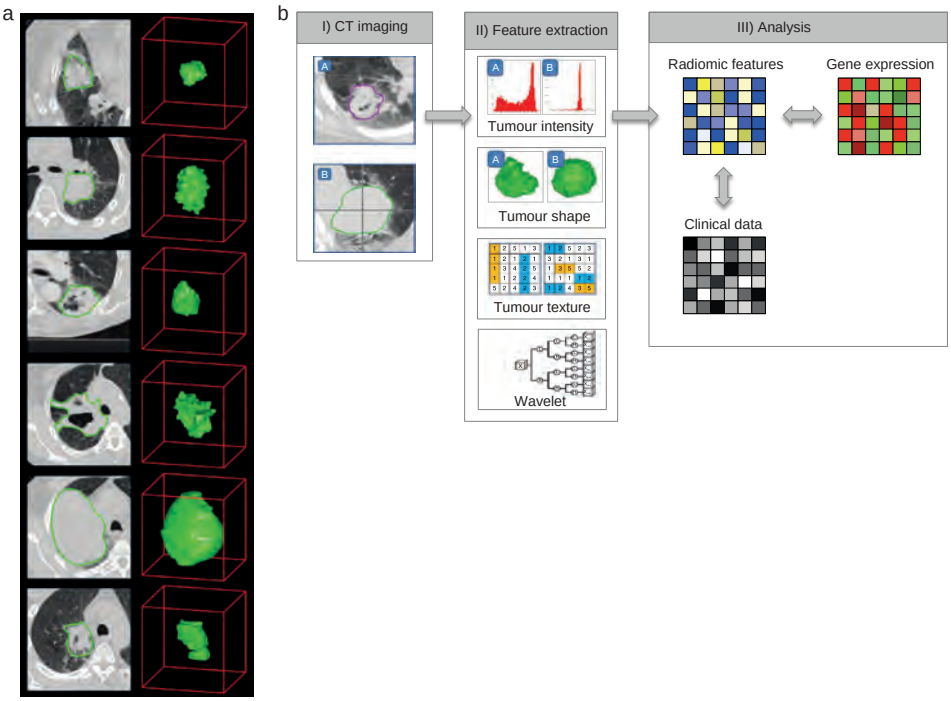


Figure 1 – Extracting radiomics data from images. (a) Tumours are different. Example computed tomography (CT) images of lung cancer patients. CT images with tumour contours left, three-dimensional visualizations right. Please note strong phenotypic differences that can be captured with routine CT imaging, such as intratumour heterogeneity and tumour shape. **(b)** Strategy for extracting radiomics data from images. (I) Experienced physicians contour the tumour areas on all CT slices. (II) Features are extracted from within the defined tumour contours on the CT images, quantifying tumour intensity, shape, texture and wavelet texture. (III) For the analysis the radiomics features are compared with clinical data and gene-expression data.

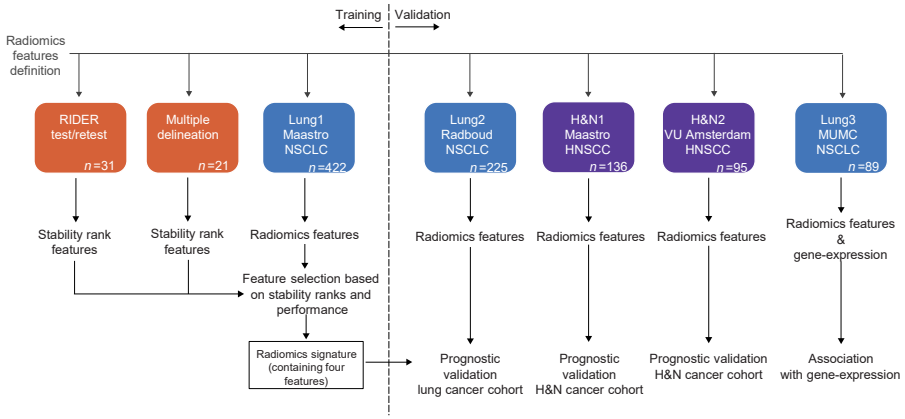


Figure 2 – Analysis workflow. The defined radiomic features algorithms were applied to seven different data sets. Two data sets were used to calculate the feature stability ranks, RIDER test/retest and multiple delineation respectively (both orange). The Lung1 data set, containing data of 422 non-small cell lung cancer (NSCLC) patients, was used as training data set. Lung2 (n = 225), H&N1 (n = 136) and H&N2 (n = 95) were used as validation data sets. The Lung3 data set (n = 89) was used for association of the radiomic signature with gene expression profiles. For the multivariate analysis, only one fixed four-feature radiomic signature was tested in the validation data sets.

RESULTS

Association of radiomic data with clinical data

To assess the value of radiomic features to capture phenotypic differences of tumours, we performed an integrated analysis assessing prognostic performance and association with gene expression in lung and head-and-neck cancer data sets. First, we defined 440 quantitative image features describing tumor phenotype characteristics by: I) tumor image intensity, II) shape, III) texture and IV) multiscale Wavelet (**Figure 1b**, **Supplementary Methods**).

To investigate radiomic expression patterns we extracted radiomic features from the Lung1 dataset, consisting of 422 non-small cell lung cancer (NSCLC) patients (**Figure 2**). Unsupervised clustering revealed clusters of patients with similar radiomic expression patterns (**Figure 3**). We compared the three main clusters of patients with clinical parameters (**Figure 3b**), and found significant association with primary tumour stage (T-stage; $p < 1\times10^{-20}, \chi^2$ test) and overall stage ($p = 3.4\times10^{-3}, \chi^2$ test), wherein cluster I was associated with lower stages. N-stage (lymph node) and M-stage (metastasis), however, showed no correspondence with the radiomic expression patterns ($p = 0.46$, and $p = 0.73$ respectively, χ^2 test).

Furthermore, a significant association with histology ($p = 0.0019, \chi^2$ test) was observed, wherein squamous cell carcinoma showed a higher presence in cluster II. Looking

at the representation of the feature groups (**Figure 3c**), there was no correspondence between the feature group and radiomic expression patterns.

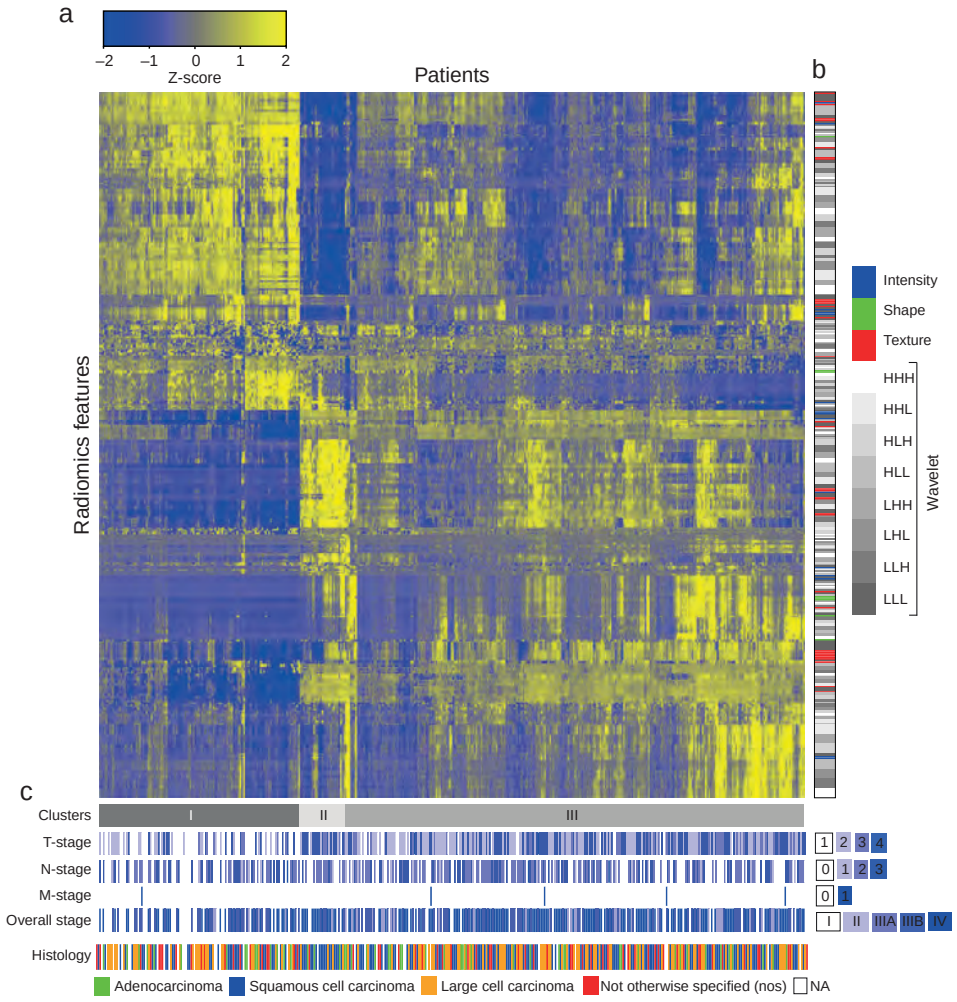


Figure 3 – Radiomics heat map. (**a**) Unsupervised clustering of lung cancer patients (Lung1 set, n=422) on the y-axis and radiomic feature expression (n=440) on the x-axis, revealed clusters of patients with similar radiomic expression patterns. (**b**) Clinical patient parameters for showing significant association of the radiomic expression patterns with primary tumour stage (T-stage; $p < 1{\times}10^{-20}, \chi^2$ test), overall stage ($p = 3.4{\times}10^{-3}, \chi^2$ test), and histology ($p = 0.0019, \chi^2$ test). (**c**) Correspondence of radiomic feature groups with the clustered expression patterns.

Prognostic value of radiomic data

The possible association of radiomic features with survival was then explored by Kaplan-Meier survival analysis. For training we used the Lung1 dataset, and for validation the Lung2, H&N1, H&N2 datasets (Figure 2). The radiomic features were not normalized on any dataset, and only the raw values were used that were directly computed from the DICOM images.

To ensure a completely independent validation, the median value of each feature was computed on the training Lung1 data set, and locked for use as a threshold in the validation data sets to assess the survival differences without retraining. In Supplementary Figure 1 we show Kaplan–Meier survival curves for four representative features. Features describing heterogeneity in the primary tumour were associated with worse survival in all four data sets. Also, patients with more compact/spherical tumours had better survival probability.

Overall, the median threshold derived from Lung1 yielded a significant survival difference for 238 features (54% of total 440; G-rho test, false discovery rate (FDR) 10%) in the Lung2 validation data set. Furthermore, there was a significant survival difference for 135 features (31%) in H&N1 and for 186 features in H&N2 (42%). Sixty-six (15%) of the features derived from Lung1 were significant for survival in all three validation data sets (Lung2, H&N1 and H&N2).

Building prognostic radiomic signature

To build a prognostic radiomic signature, the analysis was divided in training and validation phases (Figure 2). For the training phase, we first explored feature stability determined in both test-retest and inter-observer setting. Using the publicly available RIDER [10] dataset, consisting of 31 sets of test-retest CT-scans that were acquired approximately 15 minutes apart, we tested how consistent the radiomic features were between the test and retest scan. The multiple delineation dataset, where five oncologists delineated lesions on CT scans from 21 patients [11], was used to test the stability of the radiomic features to variation in manual delineations.

For each feature we compared the stability ranks for test-retest and multiple delineation with prognosis in the Lung1 training dataset. Although the stability ranks did not use any information about prognosis, in general, features with higher stability for test retest and delineation inaccuracies showed higher prognostic performance (Supplementary Figure 2). This is possibly due to reduced amount of noise in the stable features and supports the use of stability ranks for feature selection.

To test the multivariate performance of a radiomic signature, we used the workflow depicted in Figure 2 and Supplementary Figure 3. We focused our analysis on the 100 most stable features, which were determined by averaging the stability ranks of RIDER data set and multiple delineation data set. To remove redundancy within the radiomic

information, we selected the single best performing radiomic feature from each of the four-feature groups, and combined these top four features into a multivariate Cox proportional hazards regression model for prediction of survival.

The resulting radiomic signature consisted of (I) 'Statistics Energy' (**Supplementary Methods** Feature 1) describing the overall density of the tumour volume, (II) 'Shape Compactness' (Feature 16) quantifying how compact the tumour shape is, (III) 'Grey Level Nonuniformity' (Feature 48) a measure for intratumour heterogeneity and (IV) wavelet 'Grey Level Nonuniformity HLH' (Feature Group 4), also describing intratumour heterogeneity after decomposing the image in midfrequencies. The weights of each of the features in the signature were fitted on the training data set Lung1.

Prognostic validation of radiomic signature

The performance of the four feature radiomic signature was validated in the datasets Lung2, H&N1, and H&N2 (**Figure 2**) using the concordance index (CI), which is a generalization of the area under the ROC-curve [12]. The radiomic signature had good performance on the Lung2 data (CI $= 0.65, p = 2.91\times10^{-09}$, Wilcoxon test), and a high performance in H&N1 (CI $= 0.69, p = 7.99\times10^{-07}$, Wilcoxon test) and H&N2 (CI $= 0.69, p = 3.53\times10^{-06}$, Wilcoxon test). In **Figure 4a** the Kaplan–Meier curves are shown.

Although volume had a good performance in all datasets, the radiomic signature performed significantly better, suggesting that radiomic features contain relevant, complementary information for prognosis (**Supplementary Table 1**). Furthermore, combining the radiomic signature with volume was significantly better than volume alone in all datasets. Comparing the radiomic signature to the TNM staging [13], we see that the signature performance was better in both Lung2 and H&N2 and comparable in H&N1 (**Supplementary Table 1**). Importantly, combining the radiomic signature with TNM staging showed a significant improvement in all datasets, compared with TNM staging alone. Furthermore, we assessed if the radiomics signature preserved the significant prognostic performance compared to the treatment patients received. We found that the signature preserved its prognostic performance for all treatment groups (radiation, or concurrent chemo-radiation), for both Lung and H&N cancer patients (**Supplementary Table 2**), demonstrating the complementary value of radiomics for each treatment type.

Human papillomavirus (HPV) is an important determinant in head and neck cancer patients, especially those with oropharyngeal carcinoma for prognosis and may guide future treatment selection. We did not find a significant association between radiomic signature prediction and HPV status in a combined analysis in the H&N1 and H&N2 dataset ($p = 0.17$, Wilcoxon test, **Supplementary Table 3**). However, we found that the signature preserved its prognostic performance in the HPV negative group (CI $= 0.66$), consisting of the majority of patients (76%, n=130), demonstrating the complementary value of Radiomics to HPV screening.

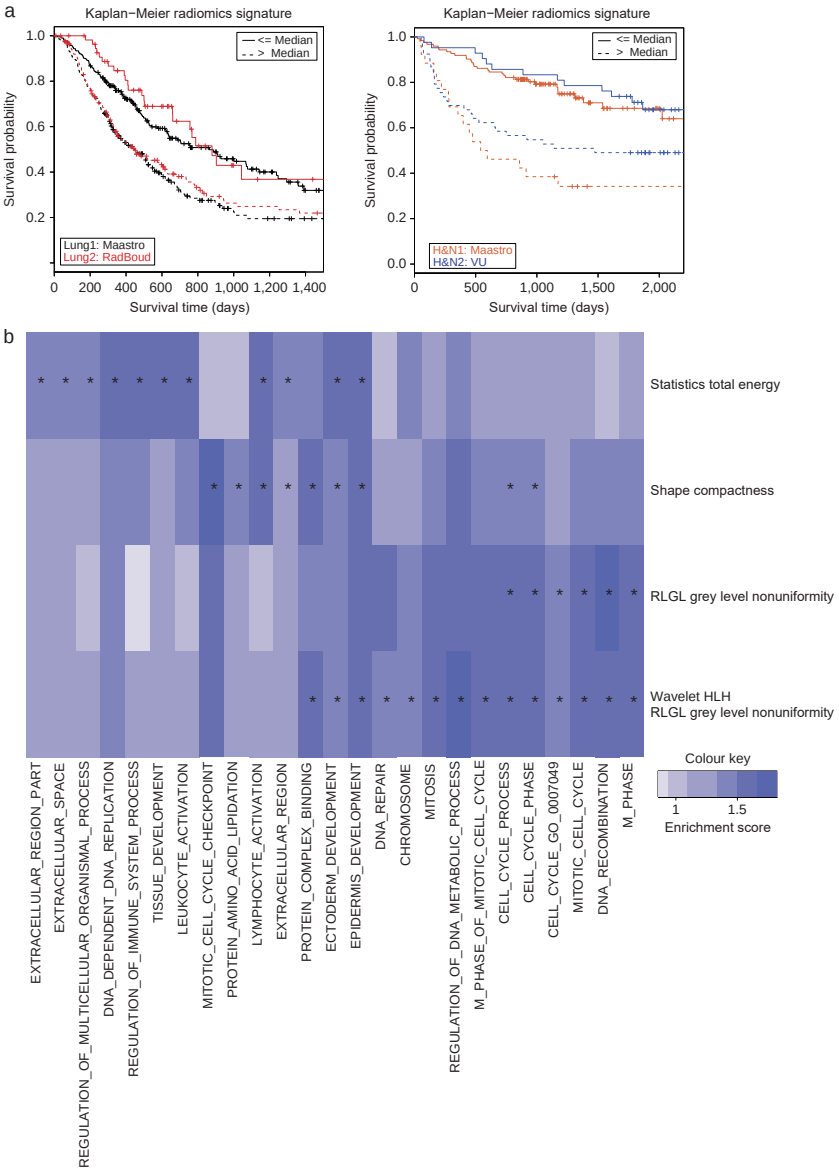


Figure 4 – Prognostic performance and gene-expression association of the radiomics signature. (a) Radiomic signature performance. Kaplan–Meier curves demonstrating performance of the radiomic signature on the lung cancer data sets (left) and the head-and-neck cancer data sets (right). The signature was built on the Lung1 data (n=422). The signature had a good performance in the Lung2 (CI $= 0.65, p = 2.91\times10^{-09}$, Wilcoxon test, n=225), and a high performance in H&N1 (CI $= 0.69, p = 7.99\times10^{-07}$, Wilcoxon test, n=136) and H&N2 (CI $= 0.69, p = 3.53\times10^{-06}$, Wilcoxon test, n=95) validation data sets. **(b)** Association of radiomic signature features and gene expression using gene-set enrichment analysis (GSEA) in the Lung3 data set (n=89). Gene sets that have been significantly enriched (FDR=20%) for at least one of the four radiomic features are indicated with an asterisk. The corresponding normalized enrichment scores (NES), GSEA's primary statistic, for all radiomic signature features is displayed in a heat map, where light blue means low and dark blue means high NES.

To assess the association between the radiomic signature and the underlying biology, we compared the radiomic signature with gene-expression profiles (Lung3 dataset, **Figure 2**) using gene-set enrichments analysis (GSEA) [1, 14]. We found significant associations between the signature features and gene-expression patterns (**Figure 4b**). Further, the radiomic features are significantly associated with different biologic gene-sets, demonstrating that radiomic features probe different biologic mechanisms. It is noteworthy that both intra-tumour heterogeneity features in the signature (Feature III and IV) were strongly correlated with cell cycling pathways, indicating an increased proliferation for more heterogeneous tumours.

DISCUSSION

Medical imaging is one of the major factors informing medical science and treatment. Its potential resides in its ability to assess the characteristics of human tissue non-invasively, and therefore is routinely used in clinical practice for oncologic diagnosis and treatment guidance and monitoring.

However, traditionally, medical imaging has been a subjective or qualitative science. Recent advances in medical imaging acquisition and analysis, allow the high-throughput extraction of informative imaging features to quantify the differences that oncologic tissues exhibit in medical imaging.

Radiomics applies advanced computational methodologies to medical imaging data, to convert medical images into quantitative descriptors of oncologic tissues [8].
In this study, we analysed 440 radiomic features quantifying tumour phenotypic differences based on its image intensity, shape and texture. In a large dataset of 1019 lung and head and neck cancer patients, of which we extracted radiomic features on computed tomography images, we found that a large number of radiomic features have prognostic power, many of which its prognostic implication have not been described before. Furthermore, our integrated analysis showed that features selected based on their stability and reproducibility were also the most informative features, which indicates the power of integrating independent datasets for radiomic feature selection and model building.

We showed as well that a radiomic signature, capturing intra-tumour heterogeneity, was strongly prognostic and validated in three independent datasets of lung and head and neck cancer patients, and was associated with gene-expression profiles. To avoid any form of over-fitting or bias, we performed a robust statistical validation: only one radiomics signature (containing 4 radiomic features) was validated in data of 545 patients in independent validation datasets (**Figure 2** and **Supplementary Figure 3**). The four features were selected based on feature stability and prognostic performance in the discovery dataset only.

The top performing feature "Gray Level Nonuniformity" (Feature 48) and the most dominant features in the radiomic signature (features III and IV), quantified intra-tumour heterogeneity. Indeed, it is often hypothesized that intra-tumour heterogeneity is exhibited on different spatial scales, for example at the radiological, macroscopic, cellular, and the molecular (genetics) level. Radiological tumour phenotype characteristics may thus be useful to investigate the underlying evolving biology. It is known that multiple subclonal populations co-exist within tumours, reflecting extensive intra-tumoral "somatic evolution" [15, 16]. This heterogeneity is a clear barrier to the goal of personalized therapy based on molecular biopsy-based assays, as the identified mutations and gene-expression does not always represent the entire population of tumour cells [17, 18]. Radiomics circumvents this by assessing the comprehensive 3D tumour bulk. The study presented here probes heterogeneity and demonstrates corresponding clinical importance in two cancer types. Furthermore, we demonstrated association of intra-tumour heterogeneity with proliferation, a general hallmark of cancer.

Overall, the lung-derived radiomic signature had better performance in head and neck compared to lung cancer. One reason could be that head and neck images were acquired with head immobilization, whereas lung images were acquired with free-breathing and are affected by patient movement or respiration, resulting in relatively more image noise. Nonetheless, our results show that the radiomic signature could be transferred from lung to head and neck cancer, which suggests that the signature identifies a general prognostic tumour phenotype.

Our method provides a non-invasive (and therefore with no risk of infection or complications that accompany tissue biopsies), fast, low cost, and repeatable way of investigating phenotypic information, potentially speeding up the development of personalized medicine. Furthermore, we show that the radiomic signature is significantly associated with the underlying gene-expression patterns, suggesting that inter-patient differences of gene-expression are large than intra-patient differences.

The clinical impact of our results are illustrated by the fact that it advances knowledge in the analysis and characterization of tumours in medical images, previously not done, and provides knowledge currently not used in the clinic. We showed the complementary performance of Radiomic features with TNM staging for prediction of outcome, which illustrates the clinical importance of our findings as TNM is routinely used in the clinic. Currently, the TNM staging system is used for risk stratification and treatment decision-making. However, the TNM staging system is primarily based on resectability of the tumour, while a larger number of NSCLC patients will receive primary treatment with radiotherapy either alone or combined with chemotherapy. Therefore, the TNM staging system is insufficient for risk stratification of this group of patients, in particular to make the decision between curative treatment (concomitant radio-chemotherapy) or palliative treatment especially in elderly patients, a growing issue in western countries. Our results show that the radiomics signature is performing better in independent cohorts than the

TNM classification. In future clinical trials this inexpensive method can be used as well for pretreatment risk stratification (e.g. high, low risk).

Furthermore, we have shown for the first time the translational capability of radiomics in two cancer types (lung and head and neck cancer). These results indicate that radiomics quantifies a general prognostic cancer phenotype that likely can broadly be applied to other cancer types. Similar observations have been made in gene-expression studies where signatures are prognostic across different diseases [19].

Analysis of image features applied to medical imaging has been a largely studied field and extensive literature exists. However, the majority of previous work describes the use of imaging features focused in the detection of small nodules in for example mammograms or chest CT/PET scans, or in the differential diagnosis of malignant versus benign nodules (Computed Aided Diagnostics). However, applications and methodologies are distinct from our study. Quantitative imaging for personalized medicine is a recent field, with a limited number of publications [12, 20-27]. The main clinical question of this research is not the diagnosis, but how to extract more useful information from the tumour phenotype that can be used for personalized medicine. Therefore, we assessed the association of radiomics with clinical factors, prognosis, and gene-expression levels, using large amounts of features and with external and independent validation cohorts of patients. The most important message in our manuscript is that there is prognostic and biologic information enclosed in routinely acquired CT imaging and was evident in two cancer types.

It is known that variability in image acquisition exists across hospitals and that this is a reality in clinical practice. However, in our analysis we used data directly generated from the scanner and the features were calculated from the RAW imaging data, without any pre-processing or normalization. As there was no correction by cohort or scanner type, this illustrates the translational potential of our results and it is a strong argument in favor of a multi-centric application of radiomics. The radiomics signature had strong prognostic power in these independent datasets generated in daily clinical practice. Furthermore, we expect that with better standardization and imaging protocols, the power of radiomics will even further improve. Among others, the Quantitative Imaging Network (QIN) of the National Institute of Health (NIH), as well as the quantitative imaging biomarker alliance (QIBA), investigates future directions, by performing phantom studies and discussing with vendor's open and standardized protocols for image acquisition [2, 3].

Due to the large availability of non-invasive imaging performed routinely in a large number of cancer patients, and the automated feature algorithms, the results of this work could stimulate further research of image-based quantitative features. Also, we presented evidence that the defined radiomic feature-metrics are platform independent, though this should be studied further, and can potentially be applied to other image modalities, such as magnetic resonance imaging (MRI), or positron emission tomography

(PET). This approach can have a large impact as imaging is routinely used in clinical practice, worldwide, in all stages of diagnoses and treatment, providing an unprecedented opportunity to improve medical decision support.

METHODS

Radiomics Features

We defined 440 radiomic image features that describe tumour characteristics and can be extracted in an automated way. The features can be divided into four groups: (I) tumour intensity, (II) shape, (III) texture and (IV) wavelet features. The first group quantified tumour intensity characteristics using first-order statistics, calculated from the histogram of all tumour voxel intensity values. Group 2 consists of features based on the shape of the tumour (for example, sphericity or compactness of the tumour). Group 3 consists of textual features that are able to quantify intratumour heterogeneity differences in the texture that is observable within the tumour volume. These features are calculated in all threedimensional directions within the tumour volume, thereby taking the spatial location of each voxel compared with the surrounding voxels into account. Group 4 calculates the intensity and textural features from wavelet decompositions of the original image, thereby focusing the features on different frequency ranges within the tumour volume (**Supplementary Figure 4**). All feature algorithms were implemented in Matlab. In the **Supplementary Methods**, the feature algorithms are described.

Datasets

We applied a radiomic analysis to seven image data sets. An overview of the data sets is presented in **Figure 2**. All research was carried out in accordance with Dutch law. The Institutional Review Boards of each of the participating centres approved the studies: Lung1, Lung3, H&N1 (Maastricht University Medical Center (MUMC+), Maastricht, The Netherlands), Lung2 (Radboud University Medical Center (RUMC), Nijmegen, The Netherlands) and H&N2 (VU University Medical Center (VUMC), Amsterdam, The Netherlands). The Multiple delineation data set is publicly available (downloaded from: www.cancerdata.org). This study was conducted according to national laws and guidelines and approved by the appropriate local trial committee at Maastricht University Medical Center (MUMC1), Maastricht, The Netherlands.

- The RIDER data set consists of 31 NSCLC patients with two CT scans acquired approximately 15 min apart [10]. We used this dataset to assess stability of the features for test retest.

- The multiple delineation data set consists of 21 NSCLC patients where the tumour volume was delineated manually on CT/PET scans by five independent oncologists [11]. We used this dataset to assess stability of the features for delineation inaccuracies.
- The Lung1 data set consists of 422 NSCLC patients that were treated at MAASTRO Clinic, The Netherlands. For these patients, CT scans, manual delineations, clinical and survival data were available. We used this data set to assess the prognostic value of the radiomic features and to build a radiomic signature.
- The Lung2 data set consists of 225 NSCLC patients that were treated at Radboud University Nijmegen Medical Centre, The Netherlands. For these patients, CT scans, manual delineations, clinical and survival data were available. We used this data set to validate the prognostic value of the radiomic features and signature in an independent NSCLC cohort.
- The H&N1 data set consists of 136 head-and-neck squamous cell carcinoma (HNSCC) patients treated at MAASTRO Clinic, The Netherlands. For these patients, CT scans, manual delineations, clinical and survival data were available. We used this data set to validate the prognostic value of the radiomic features and signature in HNSCC patients.
- The H&N2 data set consists of 95 HNSCC patients treated at the VU University Medical Center Amsterdam, The Netherlands. For these patients, CT scans, manual delineations, clinical and survival data were available. We used this data set to validate the prognostic value of the radiomic features and signature in a second cohort of HNSCC patients.
- The Lung3 data set consists of 89 NSCLC patients that were treated at MAASTRO Clinic, The Netherlands. For these patients pretreatment CT scans, tumour delineations and gene expression profiles were available. We used this data set to associate imaging features with gene-expression profiles.

In the Supplementary Methods and Supplementary Tables 4–7, further descriptions of the data sets are presented. The discovery Lung1 data set, consisting of CT images for 422 NSCLC patients, and the Lung3 data set consisting of CT images and gene-expression profiling for 89 NSCLC patients, are publicly available at The Cancer Imaging Archive, Lung1: https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics and Lung3: https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics-Genomics, as well as on www.cancerdata.org.

Sample size

To reduce any form of over-fitting or bias in the multivariate analysis, we trained on data the Lung1 data sets (n = 422), selecting the features and fixing the weights, and tested

only one signature (containing four features) in data of 545 patients in the independent validation data sets. There was no need for randomization as the patients originated from distinct groups. Patients were included in the analysis with the following criteria: confirmed primary tumour, patients underwent treatment with curative intent. Excluded from this analysis were patients receiving no or palliative treatment and patients with previous lung or head-and-neck cancer

Data Analysis

An overview of the analysis is shown in **Figure 2**. The analysis was divided in training and validation phases. For the training phase, we first explored feature stability determined in both test-retest and inter-observer setting. The RIDER and Multiple Delineation datasets were used to assess stability of the features to select the most informative features for further investigation. Using the RIDER test retest dataset, we tested the stability of the radiomic features between test and retest [10]. For each patient, we extracted the radiomic features from both scans. A stability rank was calculated for each feature, using the intra-class correlation coefficient (ICC), where a lower ICC rank corresponds to a more stable feature.

We assessed the feature stability for delineation inaccuracies using a Multiple Delineation dataset [11]. All radiomic features were computed for five delineations per patient, and a stability rank per feature was calculated using the Friedman test. The Friedman test is a non-parametric repeated measurement test for a non-Gaussian population. A rank of 1 indicated the most stable feature for delineation inaccuracies and 440 the least stable feature. All 440 radiomic features were extracted for the Lung1, Lung2, H&N1, and H&N2 datasets. The radiomic features were not normalized on any dataset, and only the raw values were used that were directly computed from the DICOM image. To explore the association of the radiomics features with survival we used Kaplan-Meier analysis in a training and validation phase. To ensure a completely independent validation, the median threshold of each feature on the Lung1 dataset was computed, and then this threshold was used in the validation datasets (Lung2, H&N1, and H&N2) to split the survival curves. We used the G-rho rank test for censored survival data to test for significant difference between the two survival curves. P-values were corrected for multiple testing using by controlling the false discovery rate (FDR) of 10%, the expected proportion of false discoveries amongst the rejected hypotheses.

To assess the multivariate performance of radiomic features we build a signature. We selected the 100 most stable features, determined by averaging the stability ranks of RIDER dataset and Multiple Delineation dataset. Next, we computed the performance in the Lung 1 dataset of each of the selected 100 features using the concordance index (CI) [12]. This measure is comparable to the Area Under the Curve (AUC) but can also be used for Cox regression analysis. From each of the four feature groups, we selected the single best performing feature for prognosis in the Lung1 dataset, and combined these top four

features into a multivariate Cox proportional hazards regression model for prediction survival. The weights of the model were fitted on the Lung1 dataset. We applied the radiomic signature to the validation datasets Lung2, H&N1, and H&N2, and performance was assessed with the CI. To calculate significance between two models we used a bootstrap approach, for 100 times we calculated the CI of both models from 100 random selected samples. The Wilcoxon test was used to assess significance.

A similar approach was used to assess if the signature had significant power, compared with random ($CI = 0.5$). We used a bootstrap approach, for 100 times we calculated the CI of the radiomics signature based on 100 random selected samples with correct outcome data, as well as on 100 random chosen samples with random outcome data. This process was repeated 100 times. The Wilcoxon test was used to assess significance, between the two distributions.

To assess the complementary effect of the signature with clinical parameters, we build a new model with the prediction of the signature as one input and the clinical parameter as the other input. The weight of the clinical parameter was fitted on the training dataset Lung1.

To assess the association of the radiomic signature with gene expression we used the Lung3 dataset. Gene expression of 89 patients was measured on Affymetrix chips with the custom chipset HuRSTA_2a520709 for 21766 genes. Expression values were normalized with the RMA algorithm5 in the Affy package in Bioconductor. For each of the four features in the radiomic signature, we calculated the Spearman rank correlation to gene expression and used the corresponding p-values to obtain a rank of genes representing high to low agreement. Each of these gene ranks were used to perform a pre-ranked version of Gene Set Enrichment Analysis (GSEA) [14] on the C5 collection of MSigDB [28][28], which contains gene sets associated with specific GO terms. We only regarded gene sets of size 15 to 500. Local false-discovery-rates were calculated on the normalized enrichment scores (NES), GSEA's primary statistic, and only gene sets enriched with an FDR of <= 20% were retained. **Figure 4B** displays gene sets that have been significantly enriched (FDR <= 20%) for at least one of four radiomic features (indicated by an asterisk). The corresponding absolute NES in all of the four features are given color-coded, where light blue means low and dark blue means high NES.

ACKNOWLEDGEMENTS

Authors acknowledge financial support from the National Institute of Health (NIH-USA U01 CA 143062-01, Radiomics of NSCLC), the CTMM framework (AIRFORCE project, grant 030-103), EU 6th and 7th framework program (METOXIA, EURECA, ARTFORCE), euroCAT (IVA Interreg - www.eurocat.info), and the Dutch Cancer Society (KWF UM 2011-5020, KWF UM 2009-4454). Authors also acknowledge financial support from the QuIC-ConCePT project (Grant Agreement No. 115151).

SUPPLEMENTARY INFORMATION

http://www.nature.com/article-assets/npg/ncomms/2014/140603/ncomms5006/extref/ncomms5006-s1.pdf

REFERENCES

[1] Kurland BF, Gerstner ER, Mountz JM, Schwartz LH, Ryan CW, Graham MM, et al. Promise and pitfalls of quantitative imaging in oncology clinical trials. Magn Reson Imaging 2012;30: 1301-1312.

[2] Buckler AJ, Bresolin, L., Dunnick, N. R., Sullivan, D. C. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. Radiology 2011;258: 906-914.

[3] Buckler AJ, Bresolin L, Dunnick NR, Sullivan DC. Quantitative imaging test approval and biomarker qualification: interrelated but distinct activities. Radiology 2011;259: 875-884.

[4] Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. Nat Rev Clin Oncol 2013;10: 27-40.

[5] Jaffe CC. Measures of Response: RECIST, WHO, and New Alternatives. Journal of Clinical Oncology 2006;24: 3245-3251.

[6] Burton A. RECIST: right time to renovate? The Lancet Oncology;8: 464-465.

[7] Birchard KR, Hoang JK, Herndon JE, Patz EF. Early changes in tumor size in patients treated for advanced stage nonsmall cell lung cancer do not correlate with survival. Cancer 2009;115: 581-586.

[8] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48: 441-446.

[9] Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging 2012;30: 1234-1248.

[10] Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, et al. Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non–Small Cell Lung Cancer. Radiology 2009;252: 263-272.

[11] van Baardwijk A, Bosmans G, Boersma L, Buijsen J, Wanders S, Hochstenbag M, et al. PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes. Int J Radiat Oncol Biol Phys 2007;68: 771-778.

[12] Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. 2001.

[13] Compton CC, Byrd DR, Garcia-Aguilar J, Kurtzman SH, Olawaiye A, Washington MK. AJCC Cancer Staging Atlas. ed.: Springer; 2012.

[14] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 2005;102: 15545-15550.

[15] Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 2010;467: 1114-1117.

[16] Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. The New England journal of medicine 2012;366: 883-892.

[17] Gerlinger M, Swanton C. How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. British journal of cancer 2010;103: 1139-1143.

[18] Kern SE. Why Your New Cancer Biomarker May Never Work: Recurrent Patterns and Remarkable Diversity in Biomarker Failures. Cancer Research 2012;72: 6097-6101.

[19] Starmans MHW, Lieuwes NG, Span PN, Haider S, Dubois L, Nguyen F, et al. Independent and functional validation of a multi-tumour-type proliferation signature. British journal of cancer 2012;107: 508-515.

[20] Nair VS, Gevaert O, Davidzon G, Napel S, Graves EE, Hoang CD, et al. Prognostic PET ¹⁸F-FDG Uptake Imaging Features Are Associated with Major Oncogenomic Alterations in Patients with Resected Non–Small Cell Lung Cancer. Cancer Research 2012;72: 3725-3734.

[21] Diehn M, Nardini C, Wang DS, McGovern S, Jayaraman M, Liang Y, et al. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. Proceedings of the National Academy of Sciences 2008;105: 5213-5218.
[22] Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nat Biotechnol 2007;25: 675-680.
[23] Tixier F, Le Rest CC, Hatt M, Albarghach N, Pradier O, Metges JP, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med 2011;52: 369-378.
[24] El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit 2009;42: 1162-1171.
[25] Ganeshan B, Panayiotou E, Burnand K, Dizdarevic S, Miles K. Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival. Eur Radiol 2012;22: 796-802.
[26] Ganeshan B, Skogen K, Pressney I, Coutroubis D, Miles K. Tumour heterogeneity in oesophageal cancer assessed by CT texture analysis: preliminary evidence of an association with tumour metabolism, stage, and survival. Clin Radiol 2012;67: 157-164.
[27] Gevaert O, Xu J, Hoang CD, Leung AN, Xu Y, Quon A, et al. Non–Small Cell Lung Cancer: Identifying Prognostic Imaging Biomarkers by Leveraging Public Gene Expression Microarray Data—Methods and Preliminary Results. Radiology 2012;264: 387-396.
[28] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27: 1739-1740.

Chapter 7

External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma

Published in: Acta Oncologica. 2015;54: 1423-1429.

External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma

Ralph T.H. Leijenaar*, Sara Carvalho*, Frank J.P. Hoebers, Hugo J.W.L. Aerts, Wouter J.C. van Elmpt, Shao Hui Huang, Biu Chan, John N. Waldron, Brian O'Sullivan, Philippe Lambin

*These authors contributed equally to this work

ABSTRACT

Background

Oropharyngeal squamous cell carcinoma (OPSCC) is one of the fastest growing disease sites of head and neck cancers. A recently described radiomic signature, based exclusively on pre-treatment CT imaging of the primary tumor volume, was found to be prognostic in independent cohorts of lung and head and neck cancer patients treated in the Netherlands. Here, we further validate this signature in a large and independent North American cohort of OPSCC patients, also considering CT artifacts.

Methods

A total of 542 OPSCC patients were included for which we determined the prognostic index (PI) of the radiomic signature. We tested the signature model fit in a cox regression and assessed model discrimination with Harrell's c-index. Kaplan-Meier survival curves between high and low signature predictions were compared with a log-rank test. Validation was performed in the complete cohort (PMH1) and in the subset of patients without (PMH2) and with (PMH3) visible CT artifacts within the delineated tumor region.

Results

We identified 267 (49%) patients without and 275 (51%) with visible CT artifacts. The calibration slope (β) on the PI in a Cox proportional hazards model was 1.27 ($H_0: \beta = 1, p = 0.152$) in the PMH1 (n = 542), 0.855 ($H_0: \beta = 1, p = 0.524$) in the PMH2 (n = 267) and 1.99 ($H_0: \beta = 1, p = 0.002$) in the PMH3 (n = 275) cohort. Harrell's c-index was 0.628 ($p = 2.72e-9$), 0.634 ($p = 2.7e-6$) and 0.647 ($p = 5.35e-6$) for the PMH1, PMH2 and PMH3 cohort, respectively. Kaplan-Meier survival curves were significantly different ($p < 0.05$) between high and low radiomic signature model predictions for all cohorts.

Conclusion

Overall, the signature validated well using all CT images as-is, demonstrating a good model fit and preservation of discrimination. Even though CT artifacts were shown to be of influence, the signature had significant prognostic power regardless if patients with CT artifacts were included.

INTRODUCTION

Accounting for approximately half a million cases annually worldwide, head and neck squamous cell carcinoma (HNSCC) is a considerable cause of mortality and morbidity, with the majority of patients having locally advanced, unresectable disease [1]. Oropharyngeal squamous cell carcinoma (OPSCC) has been one of the fastest growing disease sites for HNSCC [2].

Known prognostic factors of locally advanced HNSCC include tumor category, nodal category and human papilloma virus (HPV) status, the latter in particular related to overall survival for OPSCC patients [3-6]. Other potential prognostic factors are obtained by molecular characterization of the tumor, mostly requiring tissue extraction [7-9]. The inherent limitations of biopsies are however their invasiveness and probability of misrepresenting the entire tumor due to intra-tumor heterogeneity, since they only characterize a small portion of the tumor [10]. In contrast, medical imaging is non-invasive and able to capture the entire tumor volume, including intra-tumor heterogeneity, which could provide additional information to supplement traditional tissue biopsy [11]. Nowadays, imaging is used routinely throughout the course of treatment and therefore there is ready access to this useful information.

Radiomics is a high-throughput approach to translate medical images into mineable data by extracting a large number of quantitative features describing tumor intensity, shape, and texture [12-14]. The hypothesis being that a comprehensive and robust [15-19] quantification of imaging phenotypes provides complementary and clinically relevant information, which may lead to imaging biomarkers [20]. As shown in recent studies, quantitative imaging features have prognostic value and potential in predicting clinical outcomes or treatment monitoring in different cancer types [21-26].

Here, we focus on a recently described prognostic radiomic signature, which is based exclusively on pre-treatment CT imaging of the primary tumor volume [27]. This signature was derived from non-small cell lung cancer (NSCLC) patients and independently validated to be not only prognostic in NSCLC, but as well in two HNSCC patient cohorts, of which all patients were treated in the Netherlands. In this study, we aim to further validate the prognostic value of this radiomic signature in a large and independent North American cohort of OPSCC patients ($n = 542$), also considering the presence of CT artifacts [28].

MATERIALS AND METHODS

Patients and CT imaging

Institutional research ethics board approval was obtained and the need for written consent was waived for this retrospective study. A total of 542 patients with OPSCC, treated

with curative intent at the Princess Margaret Cancer Center (PMH) between 2005 and 2010 were included in this study. Treatment consisted of radiotherapy or concurrent chemoradiotherapy, with standard fractionated IMRT up to 70Gy. All patients underwent pre-treatment CT imaging of the head and neck on one of available CT scanners (General Electric Discovery ST; General Electric Lightspeed Plus; Toshiba Medical Systems Aquillion ONE). CT scans were acquired in helical mode with a slice thickness of 2.5 mm (General Electric) or 2 mm (Toshiba), at 120 kVp and 300 mAs tube current (variable tube current for Toshiba scans). The gross primary tumor volume (GTV) was manually delineated for each patient for treatment planning purposes (**Figure 1a**). Images were visually assessed for the presence of CT artifacts (e.g. streak artifacts due to dental fillings) within the GTV (**Figure 1b**).

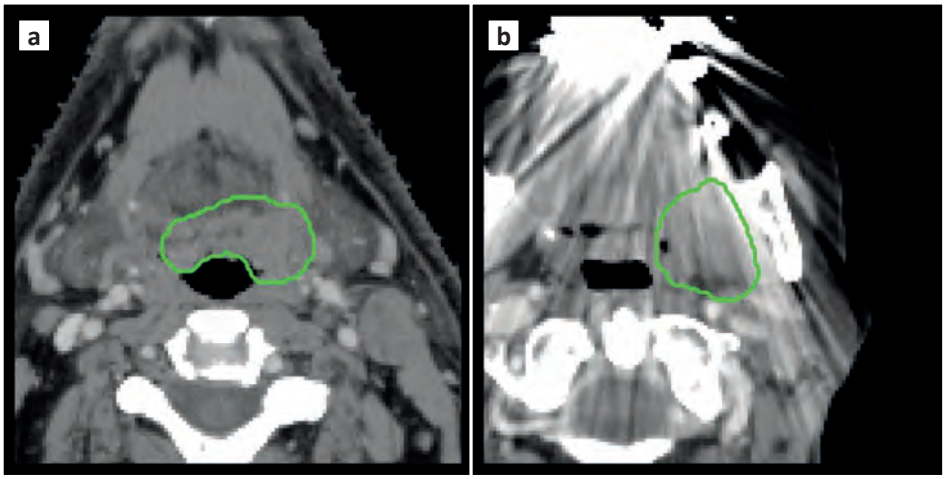


Figure 1 – Representative images of OPSCC patients from the validation cohort, without (**a**) and with (**b**) visible CT artifacts. The GTV delineation is shown in green.

Radiomic signature

The radiomic signature we aim to validate consists of the following four features, derived from the GTV: (1) "First order statistics: Energy", describing the overall density of the tumor volume; (2) "Shape: Compactness", quantifying the compactness of the tumor volume relative to that of a sphere (i.e. the most compact shape); (3) "Gray level run length: Gray level non-uniformity", a measure of intra-tumor heterogeneity; and (4) Wavelet (HLH) "Gray level run length: Gray level non-uniformity", also describing intra-tumor heterogeneity, but now after wavelet decomposition of the original CT image. A detailed mathematical description of the aforementioned features, as described by Aerts et al. [27], can be found in the Supplementary Appendix, to be found online at

http://informahealthcare.com/doi/abs/10.3109/0284186X.2015.1061214. Features were extracted using software developed in-house, in Matlab R2012b.

The radiomic signature was based on a Cox proportional hazards model and the weights (β) for each individual feature (x) in the signature are given in **Table 1**. The prognostic index (PI) for the radiomic signature, to be used for validation, is then defined as:

$$PI = \sum_i \beta_i x_i \quad (1)$$

Signature validation

To validate the radiomic signature we applied several methods, as described by Royston and Altman [29]. We first determined the model calibration slope (i.e. regression coefficient) on the PI in a Cox regression in the validation cohort and performed a likelihood ratio test of this slope being equal to 1. If the slope equals 1, the relative risk model is valid, otherwise there is a need for recalibration. We formally tested the coefficients (i.e. weights) of the individual variables of the PI, by performing a Cox regression on the individual features of the signature in the validation cohort, offsetting by the original PI (i.e. the coefficient of the PI is 1) and performing a joint test that all coefficients are 0. As a measure of model discrimination in the validation cohort, we determined Harrell's c-index, where a c-index of 1 indicates perfect discrimination. Finally we compared Kaplan-Meier survival curves between patients with a high and low signature prediction, based on a median threshold that was derived from the MAASTRO "Lung1" cohort by Aerts et al. [27]. A log-rank test was applied to test for significant differences between survival curves.

We validated the radiomic signature in the complete patient cohort (PMH1), in the subset of patients for which there were no visible CT artifacts within the delineated tumor region (PMH2) and in the subset of patients that did have visible CT artifacts (PMH3). All statistical analysis was performed in R (version 3.1.0).

RESULTS

By visual assessment, we identified all scans with CT artifacts inside the GTV, which resulted in a subset of 275 (51%) patients. In **Table 2** we summarized patient characteristic in the complete PMH validation cohort (PMH1), the patient subgroups without (PMH2) and with (PMH3) CT artifacts and, for comparison, the patient characteristics of the MAASTRO "H&N1" and VUmc "H&N2" cohorts originally used for validation of the radiomic signature by Aerts et al. [27].

In the complete PMH1 validation cohort (PMH1; n = 542), the calibration slope on the PI in a Cox proportional hazards model was found to be 1.27 ($SE = 0.175$). The slope was slightly above 1, but not significantly different from 1 ($p = 0.152$), indicating a valid

relative risk model and preservation of the discriminative value of the radiomic signature in the validation cohort. The joint test of all the predictors in the model with the PI offset was significant ($\chi^2_4 = 21.87$, $p = 2.13e-4$). Harrell's c-index for the PI was found to be 0.628 ($p = 2.72e-9$). Survival curves were significantly different ($p = 1.93e-5$) between patients with high and low radiomic signature model predictions (**Figure 2a**).

After excluding patients with visible CT artifacts within the GTV (PMH2; n = 267), the calibration slope on the PI in a Cox proportional hazards model was found to be 0.855 ($SE = 0.236$) and not significantly different from 1 ($p = 0.524$). In the model with the PI offset, the joint test of all individual feature coefficients was significant ($\chi^2_4 = 12.31$, $p = 0.015$). Harrell's c-index for the PI was 0.634 ($p = 2.7e-6$) and a significant difference between survival curves was observed ($p = 4.89e-5$) in this subset of patients (**Figure 2a**).

Considering patients with visible CT artifacts (PMH3; n = 275), the calibration slope on the PI was 1.99 ($SE = 0.273$), which was significantly different from 1 ($p = 0.002$). The joint test of all predictors was significant ($\chi^2_4 = 16.81$, $p = 0.002$) in the model with the PI offset. The c-index was found to be 0.647 ($p = 5.35e-6$) and survival curves stratified by high and low signature model predictions were significantly different ($p = 0.004$).

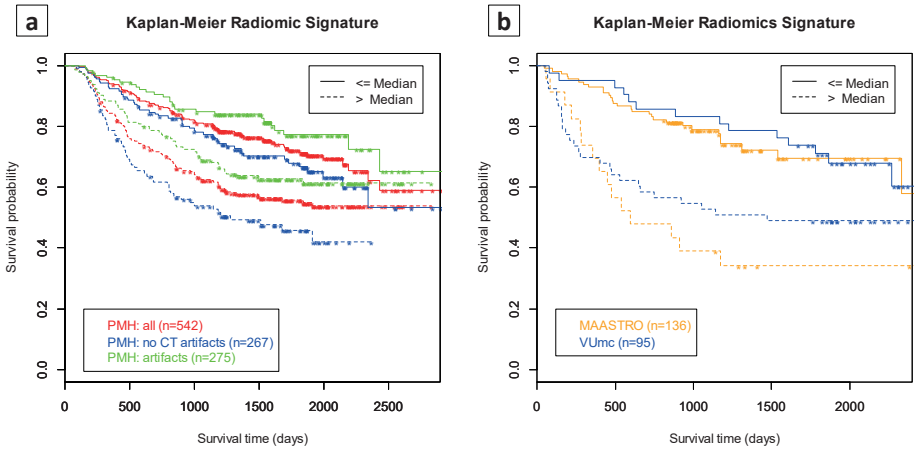


Figure 2 – Survival curves based on model predictions of the radiomic signature, split by a median prediction threshold derived by Aerts et al. from the MAASTRO "Lung1" cohort [27]. (**a**) Survival curves for the PMH validation cohort for all patients (log-rank test $p = 1.93e-5$) and for the subset of patients without (log-rank test $p = 4.89e-5$) and with (log-rank test $p = 0.004$) visible CT artifacts within the GTV. (**b**) Survival curves for the MAASTRO "H&N1" (log-rank test $p = 8.48e-05$) and VUmc "H&N2" (log-rank test $p = 0.030$) cohorts as reported by Aerts et al.

Table 1 – Description and Cox proportional hazard weights of each feature in the radiomic signature.

Feature type	Feature name	Weight
First order statistics	Energy	$2.42e-11$
Shape	Compactness	$-5.38e-03$
Gray level run length	Gray level non-uniformity	$-1.47e-04$
Wavelet (HLH) Gray level run length	Gray level non-uniformity	$9.39e-06$

Table 2 – Patient characteristics of the PMH validation cohort (PMH1), the subset of patients with no visible CT artifacts within the GTV (PMH2), the subset of patients with visible CT artifacts (PMH3), the MAASTRO "H&N1" cohorts and the VUmc "H&N2" cohort [27]. #HPV status only for oropharyngeal patients.

Variable	Frequency (%)				
	PMH1 (n=542)	PMH2 (n=267)	PMH3 (n=275)	MAASTRO (n=136)	VUmc (n=95)
Gender					
Male	79.9	82.0	77.8	81.5	65.3
Female	20.1	18.0	22.2	18.5	34.7
Primary tumor site					
Oropharynx	100.0	100.0	100.0	64.0	100.0
Larynx	0	0	0	36.0	0
T-category					
T1	12.5	13.9	11.3	25.9	10.5
T2	31.9	32.6	31.3	23.0	32.6
T3	33.4	33.3	33.5	17.8	35.8
T4	22.1	20.2	24.0	33.3	21.1
N-category					
N0	17.3	16.1	18.5	45.2	44.2
N1	10.3	10.1	10.5	11.9	11.6
N2	65.7	65.5	65.8	40.7	42.1
N3	6.6	8.2	5.1	2.2	2.1
Overall stage					
Stage I	12.5	13.9	11.3	18.5	8.4
Stage II	31.9	32.6	31.3	8.1	18.9
Stage III	33.4	33.3	33.5	17.0	18.9
Stage IVA	15.7	13.1	18.2	54.3	45.3
Stage IVB	6.5	7.1	5.8	2.2	7.4
Stage IVC	0	0	0	0	1.1
HPV/p16 status	*p16*	*p16*	*p16*	*HPV#*	*HPV*
Positive	56.3	49.4	62.9	28.4	18.9
Negative	24.0	30.3	17.8	71.6	81.1
Unknown	19.7	20.2	19.3	0	0
Treatment					
Radiotherapy	49.1	55.4	42.9	74.1	58.9
Chemoradiotherapy	50.9	44.6	57.1	25.9	41.1

DISCUSSION

An important step towards clinically using radiomics in the context of personalized medicine [14], is independent and external validation of proposed signatures [3, 4]. Here, we evaluated the validity of a recently published CT-based radiomic signature [27]. This signature was described to be prognostic in independent cohorts of both lung and head and neck cancer patients. We found that this signature validated as well in an additional large cohort of OPSCC patients.

As specified in the original publication, the radiomic features of which the signature consists, were calculated from the imaging data as-is [27]. CT images were therefore used as generated by the CT scanner and no pre-processing or normalization was performed before feature calculation. Even though it is known that scanner parameters (i.e. slice thickness or reconstruction kernels), which differ across and within patient cohorts, affect textural features computed from CT images [30], Aerts et al. [27] showed translational potential of the radiomic signature across different cohorts. This statement is further strengthened by our findings, given the good model fit and preservation of discriminative value of the signature in our validation cohort (PMH1). In comparison, Aerts et al reported c-indices of 0.686 and 0.685 in two independent head and neck cancer cohorts, whereas we found a c-index of 0.628. Furthermore, survival curves were significantly different, based on a median threshold of signature predictions, derived by Aerts et al. from the MAASTRO "Lung1" cohort. These results are in line with what has been reported by Aerts et al., for both the MAASTRO "H&N1" ($\mathrm{p} = 8.48e-05$) and VUmc "H&N2" ($\mathrm{p} = 0.030$) cohorts and a side-by-side comparison of survival curves is depicted in **Figure 2**. Even though our study endorses translational potential of the radiomic signature, we believe that standardization of imaging protocols should be pursued to eliminate variability in radiomic features between institutes, which will greatly improve the potential of radiomics [31, 32].

Another common concern in CT images of head and neck cancer are artifacts, mostly caused by metallic dental fillings or other high atomic number material implants [28]. It has to be taken into consideration that the radiomic signature was derived from CT imaging of NSCLC patients, where these type of artifacts (e.g. due to pacemakers) are uncommon. As an additional step we therefore validated the radiomic signature as well on the subset of patients without (PMH2) and with (PMH3) any visible CT artifacts within the delineated tumor region. In the PMH2 cohort subset, the calibration slope deviated less from 1 than in the complete (PMH1) cohort, signifying a better fit of the relative risk model. In contrast, the relative risk model was found to be invalid in the PMH3 cohort, indicating a need for recalibration of the model. These results suggest that there is an influence of CT artifacts on the model fit. Regardless the inclusion of patients with CT artifacts, the discriminative value of the radiomic signature was however preserved in both the PMH2 and PMH3 patient subsets, supported by Harrell's c-indices of 0.634 and 0.647, respectively. A significant difference between survival curves, stratified by high

and low signature model predictions, was preserved as well in both patient subsets (**Figure 2a**). The extent of CT artifacts and the impact on imaging features for head and neck cancer will vary between patients. Promising techniques for metal artifact reduction in CT have been reported [33]. Since radiomics relies on extracting meaningful information from medical images, techniques like these should however be thoroughly evaluated (i.e. they should not modify or introduce artificial texture). Besides the influence of the presence of CT artifacts, we also found evidence (joint tests of all the predictors in the model with the PI offset) that the fit of the radiomic signature model in our validation could be improved by adjusting the original weights of the predictors in the PI, regardless of the validity of the relative risk model.

Here we focused on the prognostic value of a radiomic signature, which only contains information derived from standard medical imaging. While our validation study provides further evidence for the concept of radiomics, we do believe that proven prognostic factors like HPV status and other clinical parameters [3-6] should as well be carefully considered in addition to radiomic information. Indeed, Aerts et al already pointed out that HPV screening for instance provides complementary information to the radiomic signature [27]. Deriving a novel and disease specific signature for head and neck cancer [34], combining radiomic and clinical information, is therefore warranted for future research–a process that should again be followed by independent validation.

CONCLUSION

We externally validated a previously described CT-based prognostic radiomic signature in a large OPSCC cohort. Overall, the signature validated well using all CT images as-is, demonstrating a good model fit and preservation of discrimination. Our results showed that CT artifacts are of influence. However, the signature had significant prognostic power regardless if patients with CT artifacts were included. Besides CT artifacts, proven prognostic factors like HPV status should as well be carefully considered, and deriving a novel and disease specific signature is warranted..

ACKNOWLEDGMENTS

We acknowledge financial support from the QuIC–ConCePT project, which is partly funded by EFPIA companies and the Innovative Medicine Initiative Joint Undertaking (IMI JU) under Grant Agreement No. 115151. This research is also supported by the Dutch technology Foundation STW (grant No. 10696 DuCAT), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs. We also acknowledge financial support from the National Institute of Health (NIH-USA U01 CA 143062-01, and NIH-USA U01 CA 190234-01), EU 7th framework program (EURECA, ARTFORCE), Kankeronderzoekfonds Limburg from the Health Foundation Limburg and the Dutch Cancer Society (KWF UM 2011 – 5020, KWF UM 2009 – 4454).

REFERENCES

[1] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin 2011;61: 69-90.

[2] Pytynia KB, Dahlstrom KR, Sturgis EM. Epidemiology of HPV-associated oropharyngeal cancer. Oral oncology 2014;50: 380-386.

[3] Rietbergen MM, Witte BI, Velazquez ER, Snijders PJ, Bloemena E, Speel EJ, et al. Different prognostic models for different patient populations: validation of a new prognostic model for patients with oropharyngeal cancer in Western Europe. British journal of cancer 2015.

[4] Rios Velazquez E, Hoebers F, Aerts HJ, Rietbergen MM, Brakenhoff RH, Leemans RC, et al. Externally validated HPV-based prognostic nomogram for oropharyngeal carcinoma patients yields more accurate predictions than TNM staging. Radiother Oncol 2014;113: 324-330.

[5] Bentzen J, Toustrup K, Eriksen JG, Primdahl H, Andersen LJ, Overgaard J. Locally advanced head and neck cancer treated with accelerated radiotherapy, the hypoxic modifier nimorazole and weekly cisplatin. Results from the DAHANCA 18 phase II study. Acta Oncol 2015: 1-7.

[6] Ampil F, Chaudhery S, Devarakonda S, Mills G. Extended survival after chemotherapy and conservative radiotherapy for HPV-16 positive stage IVB oropharyngeal carcinoma. Acta Oncol 2013;52: 1236-1237.

[7] Hoeben BA, Starmans MH, Leijenaar RT, Dubois LJ, van der Kogel AJ, Kaanders JH, et al. Systematic analysis of 18F-FDG PET and metabolism, proliferation and hypoxia markers for classification of head and neck tumors. BMC cancer 2014;14: 130.

[8] Szentkuti G, Danos K, Brauswetter D, Kiszner G, Krenacs T, Csako L, et al. Correlations Between Prognosis and Regional Biomarker Profiles in Head and Neck Squamous Cell Carcinomas. Pathology oncology research : POR 2014.

[9] De Ruyck K, Duprez F, Ferdinande L, Mbah C, Rios-Velazquez E, Hoebers F, et al. A let-7 microRNA polymorphism in the KRAS 3'-UTR is prognostic in oropharyngeal cancer. Cancer epidemiology 2014;38: 591-598.

[10] Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. The New England journal of medicine 2012;366: 883-892.

[11] Panth KM, Leijenaar RT, Carvalho S, Lieuwes NG, Yaromina A, Dubois L, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. Radiother Oncol 2015;116: 462-466.

[12] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48: 441-446.

[13] Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging 2012;30: 1234-1248.

[14] Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. Nat Rev Clin Oncol 2013;10: 27-40.

[15] Rios Velazquez E, Aerts HJ, Gu Y, Goldgof DB, De Ruysscher D, Dekker A, et al. A semiautomatic CT-based ensemble segmentation of lung tumors: comparison with oncologists' delineations and with the surgical specimen. Radiother Oncol 2012;105: 167-173.

[16] Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One 2014;9: e102107.

[17] Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol 2013;52: 1391-1397.

[18] Balagurunathan Y, Kumar V, Gu Y, Kim J, Wang H, Liu Y, et al. Test-retest reproducibility analysis of lung CT image features. Journal of digital imaging 2014;27: 805-823.

[19] Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol 2010;49: 1012-1016.
[20] Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CM, et al. 'Rapid Learning health care in oncology' - an approach towards decision support systems enabling customised radiotherapy'. Radiother Oncol 2013;109: 159-164.
[21] Rao SX, Lambregts DM, Schnerr RS, van Ommen W, van Nijnatten TJ, Martens MH, et al. Whole-liver CT texture analysis in colorectal cancer: Does the presence of liver metastases affect the texture of the remaining liver? United European gastroenterology journal 2014;2: 530-538.
[22] Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RT, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 2015;114: 345-350.
[23] Balagurunathan Y, Gu Y, Wang H, Kumar V, Grove O, Hawkins S, et al. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. Translational Oncology 2014;7: 72-87.
[24] Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, et al. Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. Int J Radiat Oncol Biol Phys 2014;90: 834-842.
[25] Ganeshan B, Panayiotou E, Burnand K, Dizdarevic S, Miles K. Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival. Eur Radiol 2012;22: 796-802.
[26] Zhang H, Graham CM, Elci O, Griswold ME, Zhang X, Khan MA, et al. Locally advanced squamous cell carcinoma of the head and neck: CT texture and histogram analysis allow independent prediction of overall survival in patients treated with induction chemotherapy. Radiology 2013;269: 801-809.
[27] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5: 4006.
[28] Purohit BS, Ailianou A, Dulguerov N, Becker CD, Ratib O, Becker M. FDG-PET/CT pitfalls in oncological head and neck imaging. Insights into imaging 2014;5: 585-602.
[29] Royston P, Altman D. External validation of a Cox prognostic model: principles and methods. BMC Medical Research Methodology 2013;13: 33.
[30] Zhao B, Tan Y, Tsai WY, Schwartz LH, Lu L. Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study. Translational Oncology 2014;7: 88-93.
[31] Buckler AJ, Bresolin, L., Dunnick, N. R., Sullivan, D. C. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. Radiology 2011;258: 906-914.
[32] Kurland BF, Gerstner ER, Mountz JM, Schwartz LH, Ryan CW, Graham MM, et al. Promise and pitfalls of quantitative imaging in oncology clinical trials. Magn Reson Imaging 2012;30: 1301-1312.
[33] Boas FE, Fleischmann D. CT artifacts: causes and reduction techniques. Imaging in Medicine 2012;4: 229-240.
[34] Parmar C, Leijenaar RT, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. Scientific reports 2015;5: 11044.

Chapter 8

Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study

Submitted work

Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study

Ralph T.H. Leijenaar*, Marta Bogowicz*, Arthur Jochems, Frank J.P. Hoebers, Frederik W.R. Wesseling, Shao Hui Huang, Biu Chan, John N. Waldron, Brian O'Sullivan, Derek Rietveld, C. Rene Leemans, Ruud H. Brakenhoff, Oliver Riesterer, Stephanie Tanadini-Lang, Matthias Guckenberger, Kristian Ikenberg, Philippe Lambin

* These authors contributed equally to this work

ABSTRACT

Introduction

HPV positive oropharyngeal cancer (OPSCC) is biologically and clinically different from HPV negative OPSCC. Here, we evaluate the use of a radiomic approach to identify the HPV status of OPSCC patients.

Methods

Four independent cohorts, with in total 778 OPSCC patients with HPV determined by p16 were collected. We randomly assigned 80% of all data for model training (N=628) and 20% for validation (N=150). On the pre-treatment CT images, 902 radiomic features were calculated from the gross tumor volume (GTV). Multivariable modeling was performed using least absolute shrinkage and selection operator (LASSO) model selection. To assess the impact of CT artifacts in predicting HPV (p16), a model was developed on all training data (M_{all}) and on the subset of training data without CT artifacts ($M_{no\ art}$). Models were validated on all validation data (V_{all}), and the subgroups with (V_{art}) and without ($V_{no\ art}$) artifacts. Kaplan-Meier survival analysis was performed to compare HPV status based on p16 and radiomic model predictions.

Results

The area under the receiver operator curve (AUC) for M_{all} and $M_{no\ art}$ ranged between 0.70-0.80 and was not significantly different for all validation datasets. There was a consistent and significant split between survival curves with HPV status determined by p16 (p=0.007; HR: 0.46), M_{all} (p=0.036; HR: 0.55) and $M_{no\ art}$ (p=0.027; HR: 0.49).

Conclusion

This study provides proof of concept that molecular information can be derived from standard medical images and shows potential for radiomics as imaging biomarker of HPV status.

INTRODUCTION

Over the last years, the incidence of OPSCC has shown a dramatic increase relative to other head and neck cancers, with a substantial proportion of OPSCC being linked to human papillomavirus (HPV) infections [1].

HPV positive OPSCC is biologically and clinically different from HPV negative OPSCC, which is often related to alcohol and tobacco consumption. HPV positive OPSCC has been shown to have superior response to radio-chemotherapy. Approximately 80% of HPV positive OPSCC patients achieve locoregional control and 5 years overall survival, in comparison to less than 50% of patients with HPV negative OPSCC and non-oropharyngeal head and neck cancers [2, 3]. This favorable outcome makes HPV positive OPSCC in particular interesting for de-escalation protocols [4].

Widely accepted methods for detection of HPV infection are in situ hybridization for viral DNA, HPV DNA or RNA PCR, and immunohistochemical investigation of the level of p16 expression, which strongly correlates with HPV infection [5].

Radiomics is a rapidly emerging field, introduced in 2012, which concerns with the high-throughput mining of large amounts of quantitative features, derived from (standard-of-care) medical imaging, for knowledge extraction [6-8] (also see: www.radiomics.world). Radiomics is in particular promising within decision support systems for precision medicine [9-11] and its potential to predict HPV status in head and neck cancer has been recognized [12]. Indeed, previous studies have reported radiologic differences between HPV positive and negative cases in terms of qualitative radiologist's readout [13] or perfusion CT [14]. Furthermore, exploratory radiomic studies have shown that heterogeneity of image-based density is potentially associated with HPV in OPSCC [15, 16].

In this multicenter study, we further investigate if a quantitative CT-based radiomic approach can objectively identify the HPV (p16) status of OPSCC, by developing and validating a radiomic signature on a large and international collection of patient data from four different institutions.

METHODS

Patients and CT imaging

Four independent cohorts, with a total of 778 OPSCC patients with HPV status determined by p16 immunohistochemistry and treated with curative intent by radiation therapy with/without concurrent chemotherapy, were collected from the Princess Margaret Cancer Center (N=427), the VU University Medical Center (N=158), the University Hospital Zürich (N=100) and MAASTRO clinic (N=93). All patients underwent pre-treatment contrast enhanced CT imaging of the head and neck. The gross primary tumor volume (GTV) was manually delineated for each patient for treatment planning purposes. Images

were visually assessed for the presence of CT artifacts (e.g. streak artifacts due to dental fillings) within the GTV. A more detailed description of acquired CT images for each of the included cohorts can be found in the **Supplementary description of CT imaging**. Institutional review board approval was obtained for each of the participating centers. Patients provided informed written consent, unless the need for written consent for this retrospective study was waived by the participating center.

Image analysis

Prior to analysis, all images were resampled to isotropic voxels of 2 mm, using linear interpolation [17]. A total of 902 radiomic features were calculated, divided into five groups: tumor intensity, shape, texture, Wavelet and Laplacian of Gaussian. All features were extracted using in-house developed software, using Matlab 2014a (MathWorks, Natick, Massachusetts, USA). Feature descriptions and mathematical definitions can be found elsewhere [8, 18]. To calculate wavelet features, we used the low pass approximation and the high pass decomposition (i.e. applying either a low or high pass filter in each direction, respectively), since these decompositions are directionally invariant. For Laplacian of Gaussian features, the texture size (fine to coarse) was highlighted by modifying the Gaussian radius parameter from 2 mm to 7 mm with 1 mm increments. Textural features were computed discretizing image intensities into bins, using both a bin width of 10 HU and 25 HU [19].

Statistical analysis

We randomly assigned 80% of all data for model training (N=628) and 20% for validation (N=150), with balanced HPV status, institution, and number of patients with visible CT artifacts.

Highly correlated features were first removed from further analysis by evaluating all pair-wise correlations. For each highly correlated feature pair ($\rho>0.9$), the variable with the largest mean absolute correlation with all remaining features was removed.

Multivariable logistic regression was performed using the least absolute shrinkage and selection operator (LASSO) model selection technique [20], with 100 times repeated 10-fold cross-validation to select the optimal tuning parameter (λ). To further reduce the chance of overfitting on the training data, we selected the simplest candidate model (i.e. the model with fewest non-zero coefficients) that is within one standard error of the best performing model. The area (AUC) under the receiver operator curve (ROC) was used to assess out-of-sample model performance in predicting HPV (p16) status.

Finally, we compared Kaplan-Meier survival curves between patients with positive and negative HPV status, based on conventional p16 immunohistochemistry and based on radiomic model HPV predictions, for all validation patients. Model class predictions were made with a probability cutoff of 0.5. Overall survival was defined as the time from

start of treatment to death as a result of any cause. A log-rank test was applied to test for significant differences between survival curves.

To assess the impact of CT artifacts, a model was also developed on the subset of patients for which there were no visible CT artifacts within the GTV. All model validation was subsequently performed on the entire validation data (V_{all}), and the subgroups of validation patients with (V_{art}) and without ($V_{no\ art}$) CT artifacts. AUC values for paired ROC curves were compared using DeLong's test [21]. Model calibration was measured by the intercept and slope of the logistic calibration curve [22].
Statistical analysis was performed in R (v. 3.3.3).

RESULTS

Patient characteristics, including HPV status, presence of CT artifacts, and follow up time are summarized in **Table 1**.

The models developed on all training data (M_{all}; 37 degrees of freedom) and on the subset of training data without CT artifacts ($M_{no\ art}$; 50 degrees of freedom), were both validated on V_{all}, $V_{no\ art}$ and V_{art}. The resulting AUC values, logistic calibration intercepts and slopes are summarized in **Table 2**. AUC values for HPV (p16) predictions made by M_{all} and $M_{no\ art}$ were not significantly different for all validation datasets. The corresponding ROC plots are shown in **Figure 1**.

Kaplan-Meier survival curves, including numbers at risk, for all validation data V_{all} are shown in **Figure 2**. For HPV determined by p16, there was a significant split between survival curves for HPV (p16) positive and negative cases (p=0.007), with a hazard ratio of 0.46 (95% CI: 0.26-0.82). For HPV (p16) predictions by M_{all} (p=0.036) and $M_{no\ art}$ (p=0.027), we observed a similar significant split between survival curves, with hazard ratios of 0.55 (95% CI: 0.31-0.97) and 0.49 (95% CI: 0.26-0.93), respectively.

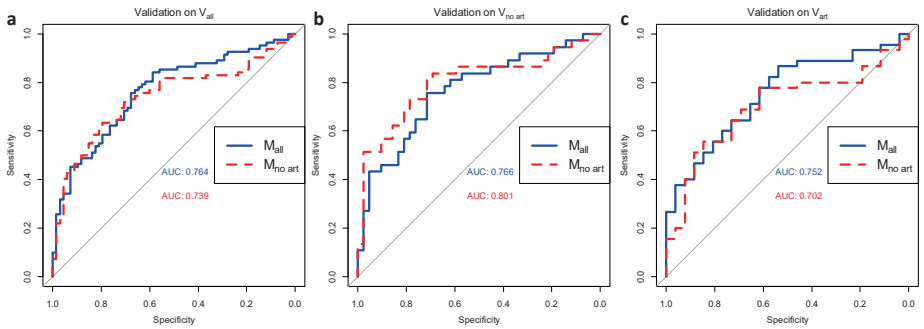


Figure 1 – ROC plots for M_{all} and $M_{no\ art}$, validated on V_{all} (a), $V_{no\ art}$ (b) and V_{art} (c).

Table 1 – HPV status, presence of CT artifacts and median follow up time for the Princess Margaret Cancer Center (PMH), the VU University Medical Center (VUmc), the University Hospital Zürich (USZ), MAASTRO clinic (MAASTRO), the 80% training data (training) and the 20% validation data (validation). For the training and validation datasets, the amount of patients from each individual cohort are given as well. Median follow up for overall survival was determined by 'reverse' Kaplan-Meier analysis (i.e. inversed censoring).

Variable	PMH (n=427)	VUmc (n=158)	USZ (n=100)	MAASTRO (n=93)	Training (n=628)	Validation (n=150)
HPV (p16) status						
Positive	303 (71%)	34 (22%)	56 (56%)	33 (35%)	344 (55%)	82 (55%)
Negative	124 (29%)	124 (78%)	44 (44%)	60 (65%)	284 (45%)	68 (45%)
CT artifacts						
Yes	219 (51%)	69 (44%)	57 (57%)	26 (28%)	300 (48%)	71 (47%)
No	208 (49%)	89 (56%)	43 (43%)	67 (72%)	328 (52%)	79 (53%)
Overall survival						
Median follow up (months)	71.6	74	44.5	51.8	69.4	65.1
Cohort						
PMH	-	-	-	-	343 (55%)	84 (56%)
VUmc	-	-	-	-	128 (20%)	30 (20%)
USZ	-	-	-	-	82 (13%)	18 (12%)
MAASTRO	-	-	-	-	75 (12%)	18 (12%)

Table 2: AUC values, logistic calibration intercepts and slopes for the model developed on all training data (M_{all}) and the model developed on the subset of training patients without CT artifacts ($M_{no\ art}$), validated in all validation data (V_{all}), the subset of validation data without CT artifacts ($V_{no\ art}$) and the subset of validation data with CT artifacts (V_{art}).

Model	Validation dataset	AUC	intercept	slope
M_{all}	V_{all}	0,7636 95% CI: 0,6874-0,8399	0,034	1,041
	$V_{no\ art}$	0,7658 95% CI: 0,6592-0,8724	-0,238	1,191
	V_{art}	0,7521 95% CI: 0,6378-0,8665	0,37	0,852
$M_{no\ art}$	V_{all}	0,7391 95% CI: 0,6582-0,8199	0,408	0,561
	$V_{no\ art}$	0,8005 95% CI: 0,6967-0,9044	0,057	1,103
	V_{art}	0,7017 95% CI: 0,5775-0,8259	0,767	0,341

DISCUSSION

In this multicenter study we developed and validated a CT based radiomic signature to predict the HPV status of OPSCC patients. In the context of radiogenomics [23, 24], our study provides a proof of concept that molecular information can be inferred from standard medical images by means of radiomics.

Previous exploratory radiomic studies that indicated a correlation between HPV infection and heterogeneity of imaging-based tumor density in OPSCC [15, 16] either were performed on small populations without validation, or only used single institution data for both model development and validation. This is a major issue in radiomic studies, as can be learned from recent literature describing the process and challenges of radiomics [8, 12, 25-27]. In this multicenter study we used a large collection of imaging data from four different institutions for model development and validation.

However, including data from different institutions introduces variety in image acquisition and reconstruction, which has been shown to affect radiomic features [28, 29]. Shafiq-ul-Hassan et al. [17] investigated voxel-size dependency of radiomic features and found that the robustness of radiomic analyses can be improved by resampling to a nominal voxel size or by normalizing the voxel size. All images in this study were therefore resampled to isotropic voxels of 2 mm, which was approximately the average slice spacing, using linear interpolation. Furthermore, as shown previously, textural features and their interpretation are affected by the bin width used to discretize image intensities [19]. Therefore, features calculated for different bin widths may provide additional predictive information. To account for this, textural features were computed using both a bin width of 10 HU and 25 HU.

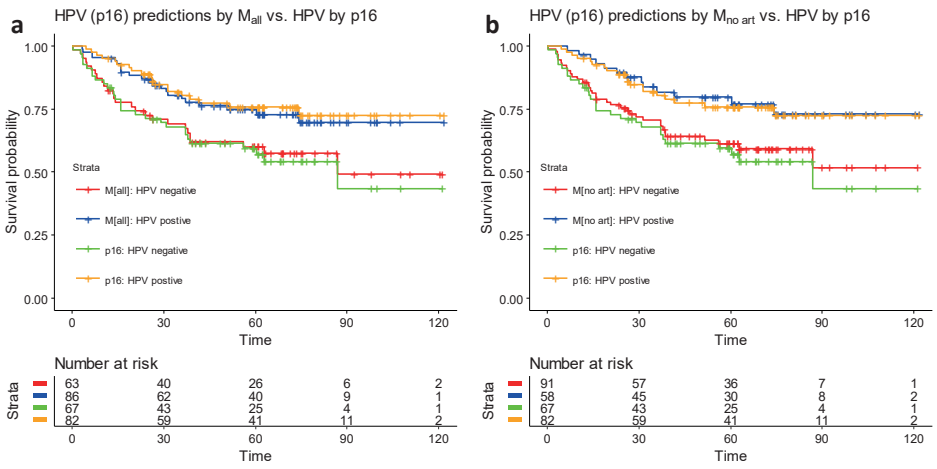


Figure 2 – Kaplan-Meier curves and number of patients at risk for HPV predictions by M_{all} vs p16 (a) and $M_{no\ art}$ vs p16 (b).

Besides variability in CT imaging, demographic differences also have to be considered. Developing a model on a single, independent cohort is therefore unlikely to sufficiently capture the variability that exists across datasets, resulting in a model with poor generalizability. We therefore performed our model development on more heterogeneous data, by randomly assigning 80% of all included data for model training and 20% for testing, with balanced HPV (p16) status, institution and number of patients with visible CT artifacts.

A common concern in the analysis of CT images of head and neck cancer are metallic dental fillings or other high atomic number material implants, which result in imaging artifacts [30]. An existing radiomic signature for overall survival [7] has previously been shown to have prognostic power regardless of CT artifacts [31]. Another recent study exploring the link between HPV status and CT radiomics, preprocessed images by completely removing artifacts affected slices from analysis [16]. However, such a process neglects potentially relevant three-dimensional information. To investigate the impact of CT artifacts on HPV prediction we developed a model on all data (M_{all}) and a model on the subset of data without artifacts ($M_{no\ art}$). What can be observed from our results is that there is no significant difference in discriminative power of both models. However, overall calibration of M_{all} was better than that of $M_{no\ art}$. It has to be noted that the extent of CT artifacts and the impact on radiomic features will vary between patients. For an individual patient, model accuracy will therefore most likely depend on the amount of the tumor region that is 'obscured' by artifacts. This would have to be further investigated, preferably including techniques for metal artifact reduction in CT.

Since HPV related OPSCC have been shown to have superior response to radio-chemotherapy [2, 3], we compared Kaplan-Meier survival curves between patients with positive and negative HPV status, based on p16 and model class predictions by M_{all} and $M_{no\ art}$ for all validation patients. Indeed, we observed a significant split ($p<0.05$) between HPV positive and HPV negative patients based on p16. For HPV (p16) predictions based on both models, we obtained survival curves similar to that of p16, with significantly different survival for HPV positive and HPV negative patients, indicating that model predictions are indeed in line with p16.

It has previously been shown that part of the OPSCC patients that test positive for p16 immunohistochemistry are in fact HPV DNA negative [32, 33]. Since HPV testing for patients included in our study was performed by p16, the likelihood of false positives has to be acknowledged. Furthermore, model class predictions (i.e. predicting either HPV positive or HPV negative), were made with a probability cutoff of 0.5, meaning that the costs for false positives and false negatives were considered equal. In clinical practice this will not be the case and false positives should be avoided (i.e. have a high cost), in order not to unjustly deescalate any patient's treatment. To achieve a clinically acceptable level of accuracy, further development and validation would be needed, including HPV DNA. It is therefore important to note that radiomic HPV prediction models are not meant to replace HPV testing and we acknowledge that clinical decision making should always be

made on the universally accepted most accurate testing (i.e. p16 and a DNA test if p16 tests positive).

Considering HPV testing is routinely performed for OPSCC patients in western countries, the clinical usefulness of a radiomic biomarker could be considered to be limited. However, our results show there is potential for radiomics to serve as a cost-effective, complementary method for HPV screening, which may also be useful in non-oropharyngeal SCCs [34]. Another potential application for a reliable radiomic biomarker could be to perform (retrospective) HPV analyses when no tissue samples are available, or in countries where it is not routinely done.

In this study, we only considered the primary tumor. However, HPV-associated OPSCC commonly present with a relatively smaller primary tumor, and relatively more advanced nodal disease. Severity of the disease may then be overestimated by the resulting higher TNM and overall stage, as these are related to other HNSCC [35]. HPV has also been shown to affect the morphology of affected lymph nodes [36]. Including radiomics of involved lymph nodes could potentially provide additional value in predicting HPV status. Furthermore, additional improvement in inferring tumor HPV status may be achieved when combining radiomics with clinical features [13].

ACKNOWLEDGEMENTS

Authors acknowledge financial support from the ERC advanced grant (ERC-ADG-2015, n° 694812 - Hypoximmuno), the QuIC-ConCePT project (IMI JU; grant no. 115151). This research is also supported by the Dutch technology Foundation STW (grant n° 10696 DuCAT & n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs. Authors also acknowledge financial support from the EU 7th framework program (ARTFORCE - n° 257144, REQUITE - n° 601826), SME Phase 2 (EU proposal 673780 – RAIL), EUROSTARS (DART), the European Program H2020-2015-17 (BD2Decide - PHC30-689715 and ImmunoSABR - n° 733008), Interreg V-A Euregio Meuse-Rhine (“Euradiomics”), Alpe d’HuZes-KWF (DESIGN), Kankeronderzoekfonds Limburg from the Health Foundation Limburg, the Zuyderland-MAASTRO grant and the Dutch Cancer Society.

REFERENCES

[1] Pytynia KB, Dahlstrom KR, Sturgis EM. Epidemiology of HPV-associated oropharyngeal cancer. Oral oncology 2014;50: 380-386.

[2] Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tan PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. The New England journal of medicine 2010;363: 24-35.

[3] Lassen P, Primdahl H, Johansen J, Kristensen CA, Andersen E, Andersen LJ, et al. Impact of HPV-associated p16-expression on radiotherapy outcome in advanced oropharynx and non-oropharynx cancer. Radiother Oncol 2014;113: 310-316.

[4] Rietbergen MM, Brakenhoff RH, Bloemena E, Witte BI, Snijders PJ, Heideman DA, et al. Human papillomavirus detection and comorbidity: critical issues in selection of patients with oropharyngeal cancer for treatment De-escalation trials. Ann Oncol 2013;24: 2740-2745.

[5] Krupar R, Hartl M, Wirsching K, Dietmaier W, Strutz J, Hofstaedter F. Comparison of HPV prevalence in HNSCC patients with regard to regional and socioeconomic factors. Eur Arch Otorhinolaryngol 2014;271: 1737-1745.

[6] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48: 441-446.

[7] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5: 4006.

[8] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol In press.

[9] Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. Nat Rev Clin Oncol 2013;10: 27-40.

[10] Lambin P, Zindler J, Vanneste B, van de Voorde L, Jacobs M, Eekers D, et al. Modern clinical research: How rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine. Acta Oncol 2015;54: 1289-1300.

[11] Lambin P, Zindler J, Vanneste BG, De Voorde LV, Eekers D, Compter I, et al. Decision support systems for personalized and participative radiation oncology. Adv Drug Deliv Rev 2016.

[12] Wong AJ, Kanwar A, Mohamed AS, Fuller CD. Radiomics in head and neck cancer: from exploration to application. Translational Cancer Research 2016;5: 371-382.

[13] Chan MW, Yu E, Bartlett E, O'Sullivan B, Su J, Waldron J, et al. Morphologic and topographic radiologic features of human papillomavirus-related and unrelated oropharyngeal carcinoma. Head Neck 2017.

[14] Nesteruk M, Lang S, Veit-Haibach P, Studer G, Stieb S, Glatz S, et al. Tumor stage, tumor site and HPV dependent correlation of perfusion CT parameters and [18F]-FDG uptake in head and neck squamous cell carcinoma. Radiother Oncol 2015;117: 125-131.

[15] Buch K, Fujita A, Li B, Kawashima Y, Qureshi MM, Sakai O. Using Texture Analysis to Determine Human Papillomavirus Status of Oropharyngeal Squamous Cell Carcinomas on CT. AJNR Am J Neuroradiol 2015;36: 1343-1348.

[16] Bogowicz M, Riesterer O, Ikenberg K, Stieb S, Moch H, Studer G, et al. CT radiomics predicts HPV status and local tumor control after definitive radiochemotherapy in head and neck squamous cell carcinoma. International Journal of Radiation Oncology*Biology*Physics 2017.

[17] Shafiq-ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys 2017;44: 1050-1062.

[18] Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RT, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 2015;114: 345-350.

[19] Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Scientific reports 2015;5: 11075.

[20] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2011;73: 273-282.
[21] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44: 837-845.
[22] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21: 128-138.
[23] Rosenstein BS, West CM, Bentzen SM, Alsner J, Andreassen CN, Azria D, et al. Radiogenomics: radiobiology enters the era of big data and team science. Int J Radiat Oncol Biol Phys 2014;89: 709-713.
[24] Panth KM, Leijenaar RT, Carvalho S, Lieuwes NG, Yaromina A, Dubois L, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. Radiother Oncol 2015;116: 462-466.
[25] Larue RT, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. Br J Radiol 2017;90: 20160665.
[26] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278: 563-577.
[27] Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present... any future? Eur J Nucl Med Mol Imaging 2017;44: 151-165.
[28] Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest Radiol 2015;50: 757-765.
[29] Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Scientific reports 2016;6: 23428.
[30] Purohit BS, Ailianou A, Dulguerov N, Becker CD, Ratib O, Becker M. FDG-PET/CT pitfalls in oncological head and neck imaging. Insights into imaging 2014;5: 585-602.
[31] Leijenaar RT, Carvalho S, Hoebers FJ, Aerts HJ, van Elmpt WJ, Huang SH, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta Oncol 2015;54: 1423-1429.
[32] Rietbergen MM, Snijders PJ, Beekzada D, Braakhuis BJ, Brink A, Heideman DA, et al. Molecular characterization of p16-immunopositive but HPV DNA-negative oropharyngeal carcinomas. Int J Cancer 2014;134: 2366-2372.
[33] Rios Velazquez E, Hoebers F, Aerts HJ, Rietbergen MM, Brakenhoff RH, Leemans RC, et al. Externally validated HPV-based prognostic nomogram for oropharyngeal carcinoma patients yields more accurate predictions than TNM staging. Radiother Oncol 2014;113: 324-330.
[34] Fujita A, Buch K, Li B, Kawashima Y, Qureshi MM, Sakai O. Difference Between HPV-Positive and HPV-Negative Non-Oropharyngeal Head and Neck Cancer: Texture Analysis Features on CT. Journal of computer assisted tomography 2016;40: 43-47.
[35] Fischer CA, Kampmann M, Zlobec I, Green E, Tornillo L, Lugli A, et al. p16 expression in oropharyngeal cancer: its impact on staging and prognosis compared with the conventional clinical staging parameters. Ann Oncol 2010;21: 1961-1966.
[36] Cantrell SC, Peck BW, Li G, Wei Q, Sturgis EM, Ginsberg LE. Differences in imaging characteristics of HPV-positive and HPV-Negative oropharyngeal cancers: a blinded matched-pair analysis. AJNR Am J Neuroradiol 2013;34: 2005-2009.

SUPPLEMENTARY DESCRIPTION OF CT IMAGING

PMH cohort

A total of 427 patients with OPSCC, treated with curative intent at the Princess Margaret Cancer Center (PMH) between 2005 and 2010 were included in this study. Treatment consisted of radiotherapy or concurrent chemoradiotherapy, with standard fractionated IMRT up to 70Gy. All patients underwent pre-treatment CT imaging of the head and neck on one of available CT scanners (General Electric Discovery ST; General Electric Lightspeed Plus; Toshiba Medical Systems Aquillion ONE). CT scans were acquired in helical mode with a slice thickness of 2.5 mm (General Electric) or 2 mm (Toshiba), at 120 kVp and 300 mAs tube current (variable tube current for Toshiba scans). In plane pixel spacing varied between 0.8 and 1.2 mm.

USZ cohort

100 stage III and IV OPSCC patients were included, treated between 2003-2015 with definitive radiotherapy (on average 70 Gy) either in combination with cisplatin (40 mg/m2, up to 7 cycles) or cetuximab (loading dose 400 mg/m2 followed by 250 mg/m2 weekly). Contrast enhanced CT images were acquired according to the institutional treatment planning protocol (tube voltage 120kV or 140kV; median tube current 214 mAs (60-450 mAs); median slice thickness 2mm (1.25mm – 3.3 mm); in-plane resolution 0.98 -1.95 mm) using three different scanners (GE Discovery STE; Siemens SOMATOM Volume Zoom; Siemens SOMATOM Definition AS). Images were reconstructed with filtered back projection algorithm.

VUmc

In total, 158 oropharyngeal squamous cell carcinoma patients, treated with curative intent at the VU Medical Center in the period 2000 – 2006 were included. Treatment options for these patients included definitive radiotherapy alone or chemo-radiation. The definitive radiotherapy regime consisted of standard fractionated radiotherapy up to 70 Gy. The concomitant chemo-radiation scheme included daily fractionation of 2Gy up to 70 Gy with a concomitant intra venous administration of cisplatin of with a dose of 100 mg/m, with or without neck dissection. All patients received a treatment planning CT scan of the head and neck (Varian Medical Systems VISION 3253). CT scans were acquired in helical mode with slice thickness of 2.5-5 mm and an in plane pixel spacing between 0.56-0.98 mm.

MAASTRO clinic

Included were 93 patients with oropharyngeal squamous cell carcinoma, treated at MAASTRO Clinic. The treatment options consisted of either definitive radiotherapy alone or concurrent chemo-radiation. Patients underwent a treatment planning 18F-FDG-PET-CT scan (Biograph, SOMATOM Sensation-16 with an ECAT ACCEL PET scanner; Siemens, Erlangen, Germany). Patients fasted at least 6 h before the start of the acquisition. A total dose dependent on the weight of the patient (weight x 4 + 20 MBq) of [18F] fluoro-2-deoxy-d-glucose (FDG), was injected intravenously, followed by physiologic saline (10 mL). Free-breathing PET and CT images were acquired after an uptake period of 45 minutes. A spiral CT (reconstructed 3 mm slice thickness and 0.98 mm in-plane pixel spacing) was performed covering the complete thoracic region.

Chapter 9

Radiomics: the bridge between medical imaging and personalized medicine

Published in: **Nature Reviews Clinical Oncology.** 2017.

Radiomics: the bridge between medical imaging and personalized medicine
Philippe Lambin, Ralph T.H. Leijenaar*, Timo M. Deist*, Jurgen Peerlings, Evelyn E.C. de Jong, Janna E. van Timmeren, Sebastian Sanduleanu, Ruben T.H.M. Larue, Aniek J.G. Even, Arthur Jochems, Yvonka van Wijk, Henry C. Woodruff, Johan van Soest, Tim Lustberg, Erik Roelofs, Wouter van Elmpt, Andre Dekker, Felix M. Mottaghy, Joachim E. Wildberger and Sean Walsh

* These authors contributed equally to this work

ABSTRACT

Radiomics is increasingly more important in cancer research; the high-throughput mining of quantitative image features from standard-of-care medical imaging that enables data to be extracted and applied within clinical decision support systems to improve diagnostic, prognostic, and predictive accuracy. Radiomic analysis exploits sophisticated image analysis tools and the rapid growth of the development and validation of medical imaging data that uses image-based signatures for precision diagnosis and treatment provides a powerful tool in modern medicine. Herein, we describe the process of radiomics, its pitfalls, challenges, opportunities, and its capacity to improve clinical decision making, emphasizing this for patients with cancer. The field of radiomics is emerging rapidly; however, this field lacks standardized evaluation of both the scientific integrity and the clinical significance of the numerous published radiomics investigations resulting from this growth. There is a clear need for rigorous evaluation criteria and reporting guidelines in order for radiomics to mature as a discipline. We therefore provide guidance through the radiomics quality score, a novel metric, together with a digital phantom (both available online) to meet this urgent need for investigations in the field of radiomics.

INTRODUCTION

Imaging is an important technology in medical science and is used in clinical practice to aid decision making [1]. The role of medical imaging however, is swiftly evolving from primarily a diagnostic tool to also include a more central role in the context of personalized precision medicine [2]. The transformation of digital medical images to mineable high-dimensional data (a field termed radiomics [3, 4]) is driven by medical images that hold information that encapsulate underlying pathophysiology [1]. This information can be harnessed through quantitative image analyses [5] and leveraged via clinical decision support systems (CDSS) [6] to improve medical decision making. Radiomics builds upon several decades of computer-aided diagnosis, prognosis, and therapeutics research [7, 8]. Radiomics is a process that identifies vast arrays of quantitative features within digital images, stores the data in (federated) databases, and subsequently mines this for knowledge extraction and application [9]. It is now possible to rapidly extract innumerable quantitative features using high-throughput computing from medical images such as computed tomography (CT), magnetic resonance (MR), and/or positron emission tomography (PET). The creation of databases that link immense volumes of radiomics data (ideally with all other pertinent data) from millions of patients to form vast rapid learning healthcare (RLHC) networks is conceivable, but presents a considerable data management hurdle [10-13].

Radiomics is not a panacea for clinical decision making. Radiomic features (such as intensity, shape, texture, wavelet, etc.) offer information on phenotype and the microenvironment that is distinct and complementary to other pertinent data sources (including clinical, treatment, and genomic data) [14]. Radiomics can augment CDSS by its combination with other pertinent data and correlating/inferencing these data sources with outcomes data, to produce accurate robust evidence-based CDSS.

The potential of radiomics to improve CDSS is beyond doubt [15] and the field is evolving rapidly. The principal challenge is optimal collection and integration of diverse multimodal data sources (imaging, clinical, treatment, genetic) in a quantitative manner that delivers definite clinical predictions that accurately and robustly predict outcomes as a function of the impending decisions [16]. Many published prediction models are available that account for factors related to both disease and treatment, but these models lack standardized evaluation of their performance, reproducibility, and/or clinical utility [17]. Consequently, these models might not be appropriate for CDSS.

In this review, we describe the process of radiomics along with recent developments in the field. The pitfalls, challenges, and opportunities presented by radiomics to improve CDSS for personalized precision oncology are highlighted, with an emphasis on the methodological aspects of radiomics prediction model development and validation. We explore the advanced and innovative information technologies that are essential for the data management of diverse multimodal data sources. Finally, we offer a vision of the

necessary steps to ensure continued progression and widespread acceptance of both radiomics and CDSS.

THE WORKFLOW OF RADIOMICS

Radiomics is defined as the quantitative mapping, that is, extraction, analysis and modelling of many medical image features in relation to prediction targets, such as clinical end points and genomics. A radiomics study can be structured in five phases: data selection, medical imaging, feature extraction, exploratory analysis, and modelling (**Figure 1**). To assess the quality of radiomics studies, we propose the radiomics quality score (RQS).

Data selection

A radiomics analysis begins with a choice of imaging protocol, the volume of interest (VOI) and a prediction target — the event one wishes to predict. Typically, the entire primary tumour is analyzed and linked to treatment outcomes, such as survival. The radiomics analysis can be performed on sub-regions of the tumour (habitats), metastatic lesions, as well as normal tissues. Analysis of these regions might yield radiosensitive phenotypes, which has implications for treatment planning strategies. Radiomics analysis, however, is not restricted to radiotherapy and can be applied to any medical image (**Figure 2**).

The importance of standardized imaging protocols to eliminate unnecessary confounding variability is recognized [9, 18]; however, non-standardized imaging protocols are commonplace. Therefore, reproducibility and comparability of radiomics studies will be achieved only by extensive disclosure of imaging protocols.

Medical imaging

Segmentation

Volumes of interest (VOIs) are segmented manually or (semi-)automatically [19]. Variability in segmentation can bias derived radiomics features [20] as segmentation determines which voxels within an image are analyzed. Multiple-segmentation is a method to limit the impact of this variation/bias. Examples to produce robust features [21] are, use multiple physicians, perturb segmentations with noise, use diverse algorithms, use different stages of the breathing cycle, etc. Key considerations are how the segmentation was done and how sensitive the radiomics analysis is to different segmentation methods [22].

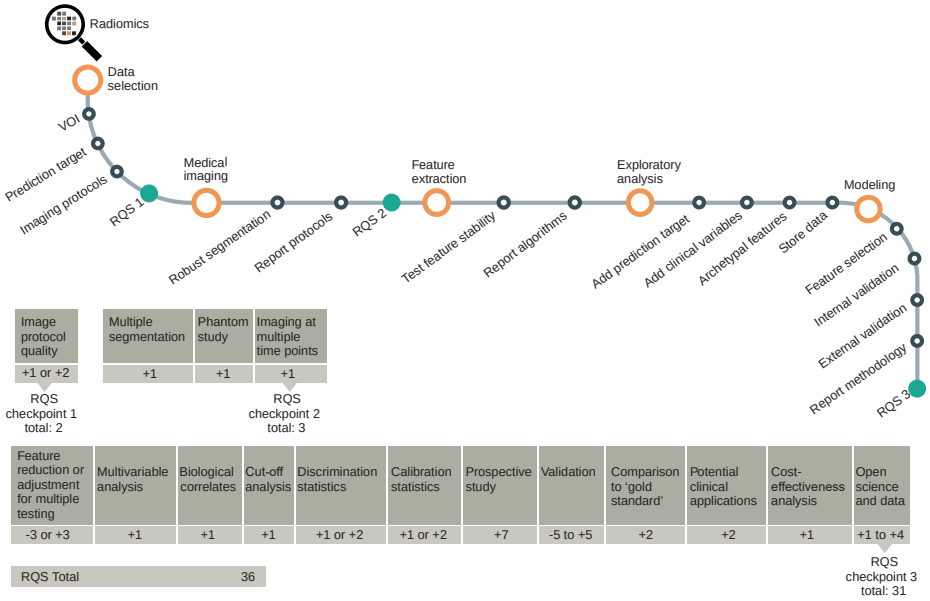


Figure 1 – Flowchart depicting the workflow of radiomics and the application of the RQS. The workflow consists of necessary steps in a radiomics analysis. The RQS both rewards and penalizes the methodology and analyses of a study, consequently encouraging best scientific practice.

Phantom studies

Investigating inter-scanner and inter-vendor variability of features is important in radiomics [23]. In cases where radiomics studies rely on data from multiple scanners, neglecting this variability can jeopardize the analysis of studies — that is, the proposed radiomics prediction model might not perform adequately on external datasets if the new data are acquired on different scanners. As data from patients scanned on multiple devices is scarce and subject to uncertainties, including organ motion, different imaging protocols, etc., phantom studies are a suitable means to gauge these uncertainties and identify features that are vendor dependent. In essence, phantom studies offer a risk mitigation strategy to help navigate from the current clinical imaging scenario to the desired optimal imaging scenario.

Imaging at multiple time points

Additional sources of variability in radiomics features are organ motion or expansion or shrinkage. Radiomics features that are strongly dependent on these factors can have limited applicability. To account for these sources of variability, available test-retest data

[24-26] can be exploited to measure radiomics feature stability. For example, two datasets of images acquired within a small period of time from a single patient cohort.

Feature extraction

The essence of radiomics is the high-throughput extraction of quantitative image features to characterize VOIs. Feature values are dependent upon factors that can include image pre-processing (filtration, intensity discretization, etc.) and reconstruction (filtered back projection, iterative reconstruction, etc.). Furthermore, there is variation in feature nomenclature, mathematical definition, methodology and software implementation of the applied feature extraction algorithms [27-29]. In order to facilitate interoperability of radiomic features, differences in nomenclature, algorithms, software implementations, as well as other methodological aspects must be elucidated.

Exploratory analysis

Radiomics and non-radiomics features together with the prediction target should be combined to create a single dataset. This enables the investigation of relationships between features. Groups of highly correlated radiomics features can be identified via clustering and these features can be reduced to single archetypal features per cluster. Radiomics features that are well correlated with routine clinical features (such as tumour stage) do not provide additional information. Auxiliary feature data collected from multiple segmentation, multiple imaging, and phantom studies, can be exploited to assess feature robustness. Volatile or robust features can be identified and subsequently excluded from model development. For example, a feature that is robust for the prediction of overall survival for lung cancer (that is imaged and segmented in a certain way) for a given dataset could be volatile for the prediction of pneumonitis in lung cancer (imaged and segmented in a certain way) for a given dataset. Thus, the process of feature reduction and exclusion should be described clearly.

Modelling

Radiomics modelling consists of three major aspects: feature selection, modelling methodology, and validation. Feature selection should be data-driven owing to the vast inhuman range of possible radiomics features; such analysis should be performed in a robust and transparent manner. To achieve holistic models, features beyond radiomics (such as clinical, treatment, biological/genetic) should also be selected. With respect to modelling methodology, identification of optimal machine-learning methods for radiomic applications is a crucial step towards stable and clinically relevant CDSS; thus in the ideal scenario multiple machine-learning methods should be employed [30] and the implementation comprehensively documented. An unvalidated model is of limited value. Validation is an

indispensable component of a complete radiomic analysis. Models must be internally validated and should be externally validated.

Feature selection

Depending on the number of filters, feature categories, and other adjustable parameters, the possible number of radiomics features that can be extracted from images is virtually unlimited. Including all possible features in a model will inevitably result in overfitting, which jeopardizes model performance in unseen patients. To avoid overfitting, features that lack robustness against sources of variability should be eliminated and archetypal features selected via dimensionality reduction techniques (for example, principal component analysis or clustering). For example, a feature that is archetypal for the prediction of overall survival in patients with lung cancer for a given dataset (imaged and segmented in a certain way) could be redundant for the prediction of pneumonitis in lung cancer for a given dataset (imaged and segmented in a certain way).

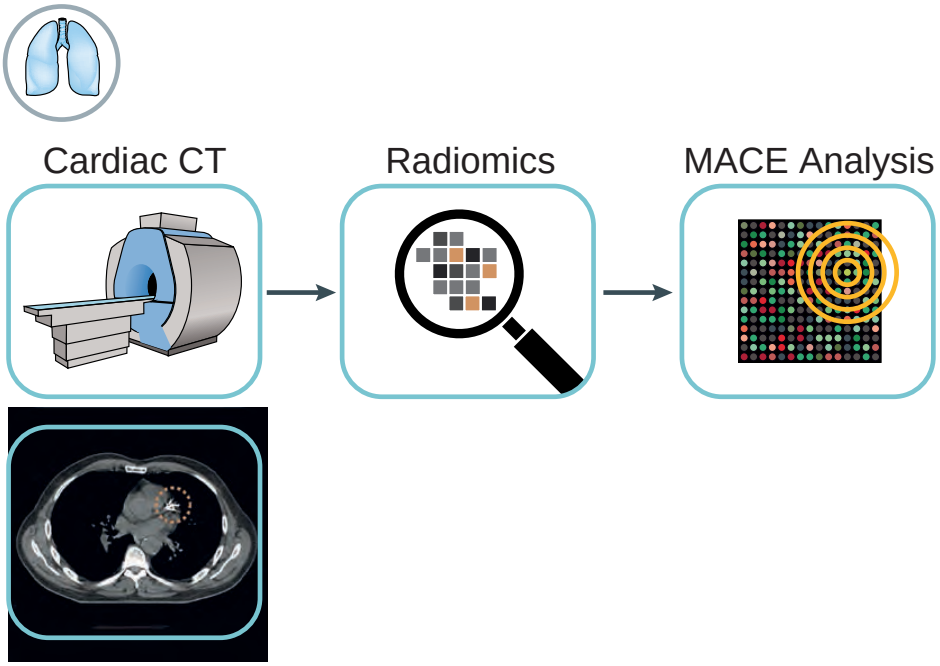


Figure 2 – Radiomics in cardiology: the current gold standard for quantification of coronary calcifications visible on CT is the 'Agatston' method (based upon intensity and volume). Radiomic features may improve quantification, differentiation between calcified and non-calcified plaque, and thus the prediction of Major Adverse Cardiac Events (MACE).

Modelling methodology

The choice of modelling methodology is often a single technique dependent on the preference and experience of those conducting the study. Different techniques are associated with different inherent limitations; these include, the independence assumption for features in logistic regression, the need for feature discretization in Bayesian networks, or the network configuration dependency in deep learning. The choice of modelling technique has been shown to affect prediction performance in radiomics [30]. Thus, multiple modelling methodology implementations are desirable, but not essential. The key aspect is that the implemented modelling methodology is reported in such a way as to make the work entirely reproducible. Ideally, by making the software code available (for example, via github [31]). (See supplementary material for an overview of machine learning techniques).

Validation

Validation (internal and/or preferably external) must be performed. Validation techniques are useful tools to assess model performance; in other words, is the model predictive for the actual population or just this subset of samples analyzed? Model performance is typically measured in terms of discrimination and calibration. Discrimination can be reported in terms of the receiver operating characteristic (ROC) curve or the area under the ROC curve (AUC). The AUC quantifies the sensitivity and specificity of the model and is equal to the probability that a randomly selected patient with the outcome is assigned a larger event-probability by the prediction model than a randomly chosen patient without the outcome. Calibration refers to the agreement between observed outcomes and model predictions, typically based on grouping of predictions. For example, the predictions are grouped into high-medium-low probability. If the mean prediction of tumour recurrence in the high probability group is 25%, the observed frequency of tumour recurrence in this group should ideally be 25 out of 100 patients. Calibration can be reported in a calibration plot and by reporting Calibration-in-the-large/slope. A measure of overall performance is the Brier score, the mean squared prediction error. All statistics should be reported for training data and validation data. Valid models should exhibit consistency in these statistics between the training and validation sets. Bootstrapping techniques can be used to estimate confidence intervals for the abovementioned statistics and should be reported. An externally validated model has much more credibility than an internally validated model as the data are considered more independent and thereby the validation. There is abundant literature available on validation techniques [32-35].

Reporting open science and data

Validation is the first step towards a model being accepted in both the scientific and clinical communities. Independent verification of the results is a necessary additional step. Reproduction means verification of the results by independent researchers repeating the

analysis using the same technique and the same data/patients, ensuring that the analysis was conducted without error. Replication means independent verification of the results by independent researchers repeating the analysis using the same technique and different (but appropriate) data/patients, leading to a stronger affirmation of the findings [36-39]. Radiomics studies involve multiple complex sub-processes (such as data selection, image acquisition, feature extraction, modelling), each shaped by a wide range of decisions, non-standardized terminology, parametrizations, and software. Reproducibility and replicability in radiomics is impossible if researchers do not disclose these intricacies. The amount of necessary information far exceeds the limits of a traditional manuscript. We propose that future radiomics publications should provide the following as supplementary material: imaging protocols, the analyzed scans, segmentations of VOIs, detailed accounts of how features are extracted including the formulae, and the modelling methodology (ideally, the code). This level of meticulous detail is required in order to facilitate reproduction and replication. Furthermore, multiple radiomics software packages are available and are subject to updates or version-control. We recognize that publishing patient data might not be possible in all circumstances. As a minimal means of comparison and to alleviate this lack of transparency, we propose that researchers publish numerical values of their investigated features computed on the digital phantom described in the supplementary material of this manuscript (available online [40]).

To compare different software implementations for radiomic feature extraction algorithms, we provide CT data of the primary tumour region and the corresponding tumour contours of four lung cancer cases, to serve as 'real life' digital phantoms (**Figure 3**). (See supplementary material for a detailed description of the digital phantom image data).

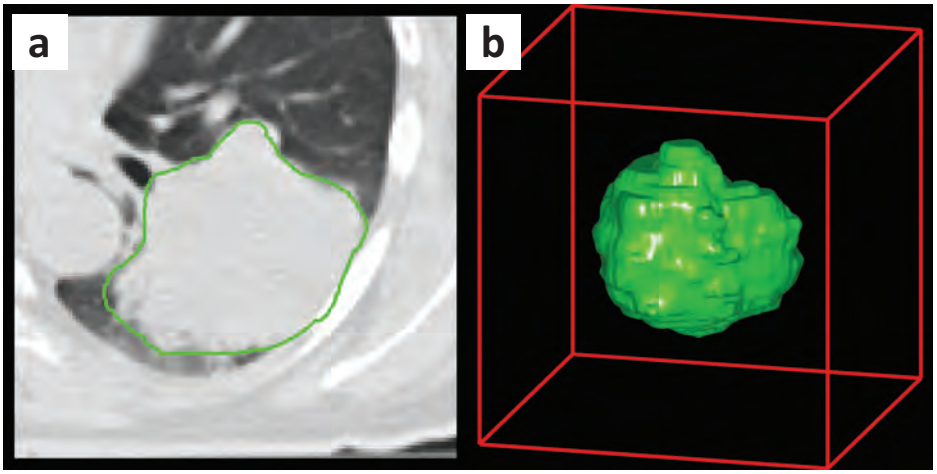


Figure 3 – Radiomics digital phantom data: (a) Representative image of a digital phantom CT image, with the tumor delineation shown outlined in green. (b) A 3D rendering of the tumor region. (please see the supplementary material).

THE RADIOMICS QUALITY SCORE

There is an urgent need for evaluation criteria and reporting guidelines in order for radiomics to develop as a field. We propose the radiomics quality score (RQS) [41] to aid assessment of both past and future radiomic studies.

Editors, reviewers, and readers should be able to easily ascertain if a radiomic study is compliant with best-practice, or alternatively if study investigators have sufficiently justified any non-compliance with guidelines. Publications should clearly state how the study has advanced the field of radiomics by specifically identifying an exigent unmet need. Overly optimistic claims concerning robustness and generalizability diminish scientific and clinical impact and should be avoided. Publications should extensively report study-design, protocols, detailed quality assurance processes, and standard operating procedures. Although the minute technical details of radiomics are tedious, they can greatly influence robustness, generalizability, and confound meta-analyses. Rigorous reporting guidelines are necessary for radiomics to mature [42-44]. Many journals now encourage and facilitate extensive supplementary materials.

The criteria of the RQS

Overwhelming evidence shows that the quality of reporting of prediction model studies is poor [32]. Full and clear reporting of information on all aspects of a prediction model to minimize bias and enhance the usefulness of prediction models is required. An excellent example is the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Initiative [32]. This initiative developed a set of recommendations for the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. We have emulated this approach in a radiomics specific context and suggest that studies should be assessed via the RQS (available online [41]), for which we identified sixteen key components; each assigned a number of points corresponding to the importance of the respective component detailed in Table 1.

RADIOMICS IN PRACTICE AND BIOLOGY LINK

Since the beginning of this decade, radiomics research has advanced dramatically revealing the potential of radiomics to substantially aid clinical care. Table 2 provides a summary of some highlighted publications to illustrate the current status of radiomics. Advances in hardware and software have enabled the realization of clinically feasible quantitative imaging of tissue pathophysiology.

Radiogenomics

Radiogenomics is associated with two closely related but distinct definitions within science, one is the study of the link between germline genotypic variations and the large clinical variability observed in response to radiation therapy [45], and the other is the study of the link between specific imaging traits and specific gene expression patterns that inform the underlying cellular pathophysiology [46]. Within the radiobiology community, a common hypothesis is that a proportion of the variance in the phenotype of interest (for example, radiation toxicity) can be attributed to genotypic variation. This hypothesis results in the clinical consequence that the risk of severe toxicity in a minority of patients limits the potentially curative doses prescribed to the majority. We posit that the overall goal of radiogenomics is to isolate the alleles and corresponding radiomic features that underlie the inherited dissimilarities in phenotype. Gene-expression profiling of various human tissues has enriched our understanding of cellular pathways and numerous pathological conditions. Investigation of different cancerous tissues in relation to samples of nonmalignant healthy tissue has elucidated tumorigenic processes and assisted in enhanced staging and sub-classification of various malignancies. Gene-expression signatures, each comprised of dozens to hundreds of genes, can significantly improve diagnosis, prognosis, and prediction of treatment response [47-52]. Seminal radiogenomic investigations showed the link between radiomic features and gene-expression patterns in patients with cancer [53-55]. One study leveraging survival data in public gene-expression data sets for patients with non-small-cell lung cancer was able to identify prognostic imaging biomarkers [54]. This radiogenomics strategy for identifying imaging biomarkers might enable a more-rapid evaluation of novel imaging modalities, thereby accelerating their translation to personalized medicine. Another investigation compared clinician-defined features extracted from contrast-enhanced CT images in patients with hepatocellular carcinoma to gene-expression patterns using machine learning with a neural network [55]. Reported combinations of 28 features could reconstruct 78% of the global gene- expression profiles associated with cell proliferation, liver synthetic function, and prognosis. In one study [53], features extracted from MR images to predict global gene-expression patterns in patients with glioblastoma multiforme revealed that an infiltrative phenotype was associated with significantly reduced survival. However, gene-expression profiling relies on surgical procurement of sampled tissue, bringing multiple risks and potential complications, and consequently rendering it unfeasible for many cancer patients. In stark contrast to genetic profiling studies, radiomic features [53, 55-59], which capture intratumoural heterogeneity in a non-invasive three-dimensional manner, can be obtained as part of routine clinical care. An example of this is the approximately 15% of globally diagnosed breast cancers designated as ER-, PR- and Her2/neu-negative, associated with a poor treatment outcome [60]. Reliable techniques for assessment of Her2Neu (FISH) are expensive and time consuming. In one study considering the tumor as well as its surrounding parenchyma on DCE-MRI radiomic image phenotyping provides

useful information for identifying these triple-negative breast cancers [61]. Presently, the radiogenomics landscape is rapidly evolving from an effort to screen a limited number of candidate genes toward an open discovery approach in the powerful, but challenging, era of RLHC [62-66].

Radiogenomics features provide valuable biomarkers for CDSS [67-70]; these include prognostic and predictive factors for outcomes [71], such as tumour response and normal-tissue tolerance. Notwithstanding these virtues, trials of radiogenomics biomarkers are susceptible to experimental and imaging inconsistency; therefore, standardization of assay criteria, image acquisition, segmentation, trial design as well as analysis is vital if radiogenomics biomarkers are to be effective diagnostic, prognostic, and predictive tools in oncology [72].

Radiosensitivity and the tumour habitat

Tumour control following radiotherapy is chiefly governed by the following criteria: the quantity of cancer stem cells (which is characteristically associated with the tumour volume that precedes therapy), the innate radiosensitivity of the stem cells, the hypoxic fraction, reoxygenation of the tumour vicinity and the repopulation capacity throughout the course of therapy [73-75].

Radiogenomic analysis of tumour habitats has the capacity to unlock knowledge with respect to these criteria [76, 77]. An example being the strong correlation reported between microvascular density and PET/MR-derived radiomics features for primary clear-cell-renal-cell-carcinoma [78]. Tumours in general display considerable variability in radiosensitivity, including those cells in the tumour of analogous origin and histological type [79-81]. In other words, relatively homogeneous tumour areas can still have high variability in radiosensitivity. Efforts to quantify the radiosensitivity of human tumours are presently founded upon the *ex-vivo* tumour survival fraction, and the detection of unrepaired DNA double strand breaks, for example, by assessment of phosphorylated histone γH2AX [82, 83]. Preclinical (prostate, lung, and brain cancers) and clinical (cervical, head-and-neck cancers) studies have proven that tumour cell radiosensitivity is a key feature for radiotherapy outcome in prostate, lung, brain, cervical, and head-and-neck cancers [84-88] . Colony assays that these data are built on, however, suffer from technical deficiencies that include poor platting efficiency (<70%) for human tumours and the protracted time required to produce data, which can be up to several weeks.

Overall, the weakness of these approaches has been the substantial experimental variability rather than interpatient variations in radiosensitivity. Non- malignant tissue toxicity is the dose-limiting factor in radiation oncology; therefore, a comprehensive CDSS should be built upon predictors of dose, tumour-control/non-malignant-tissue-complication probability, as well as cost-effectiveness [89], in order to facilitate improved escalated or de-escalated individualized treatments for patients.

Immunotherapy

In the field of oncology, a promising research area is that of biomarkers for immunotherapy, as well as imaging biomarkers [90, 91]. It is well established that radiotherapy stimulates an antitumour immune response and that the clinical potential of these immunogenic effects in concert with diverse immunotherapies is substantial[92]. For a robust and vigorous immune response, a prerequisite is the activation of T-cells coupled with antigen specificity and memory effects [93]. In this response, the tumour must express antigens that are distinct from healthy tissue, in order to be identifiable to the immune system. Such neoantigens arise from mutated proteins within the tumour cell. Research results demonstrate that the success rate of immunotherapies is dependent on the presence of neoantigen-specific T-cells [94]. Moreover, the mutational load of many types of human tumours correlates with the cytolytic activity of natural killer cells and T-cells[95]. When a tumour has the potential to be identified by the immune system, a suitable immune response can be coordinated. It is of critical importance that neoantigens are taken up by antigen-presenting cells (APCs) or dendritic cells (DCs) and subsequently cross-presented to naïve T-cells [96]. The T cells are then converted into tumour killing cytotoxic T-cells. The capacity of radiation to boost immune responses seems to be crucially dependent on the quantity and quality of DCs present in the tumour local environment [97, 98].

Tumours have diverse means to shield them from ample cytotoxic T-cell responses; for example, by interfering with several immune checkpoints. Depending on the type of interference a number of specific molecules have biomarker potential. One example is the expression of PD-1L. PD-1L binds its receptor, PD-1 on the cytotoxic T-cell, preventing cytolytic action. The mutational load of the tumour has been correlated with clinical responses to anti-PD-1-mediated immunotherapy, for example, in patients with non-small-lung cancer [99, 100]. Such biological and genetic features hold potential and radio-genomic analysis will undoubtedly have an important role in leveraging this information in future CDSS. The basic hypothesis, still to be tested, is that tumours with high mutational loads will have more neoantigens, and consequently be more heterogeneous on radiomics analysis and would be more sensitive to immunotherapy. The exact opposite scenario to that found in radio(chemo)therapy where tumour heterogeneity through radiomics analysis is an adverse prognostic factor. Regarding the combination of radio(chemo)therapy and immunotherapy, it is uncertain which effect will dominate.

Technical aspects of radiomics investigation

Accredited radio-genomic centres should be established. Stakeholders in the academic, clinical, industrial, and regulatory spheres must collaborate to create, sustain, and standardize the required best-practice framework. Radiomic studies are difficult to perform consistently, and thus, accreditation is vital to the advancement of radiomics. Techniques

for the workflow of radiomics ought to be independent of vendors and upgrades to hardware and/or software. Radiomic studies should quantify reproducibility owing to the beneficial ethical, economic and logistical effects (such as informing power calculations and required samples sizes, trial duration and trial cost). Optimal reproducibility and stability enables multicentre studies to maximize the likelihood of a validated radiomic signature being designated fit-for-purpose in routine clinical use. Prospective studies relating radiomics to clinical outcomes in appropriate patient populations and sufficiently powered are pivotal. Numerous studies are underpowered for sensitivity and specificity; however, study populations should not be skewed by selecting only those patients who are more capable of adhering to complex imaging protocols than the general population. All findings should be published, including true-negatives, false-negatives and false-positives, and the perceived adversity to negative results tempered because substantial bias risks distorting the radiomics landscape.

Economic elements of radiomics investigation

Multicentre, collaborative and federated efforts are required to share, store, and curate data. Data-sharing enables highly powered prospective studies and accelerates the development and validation of radiomic signatures derived from new and existing data. Networked centres can quickly recruit sufficient patient numbers to drive discovery and innovation. Outcome studies should include health economic considerations. Moreover, cost-per-quality-adjusted-life-year comparisons should be conducted with and without radiomics to more accurately determine the economic potential [101].

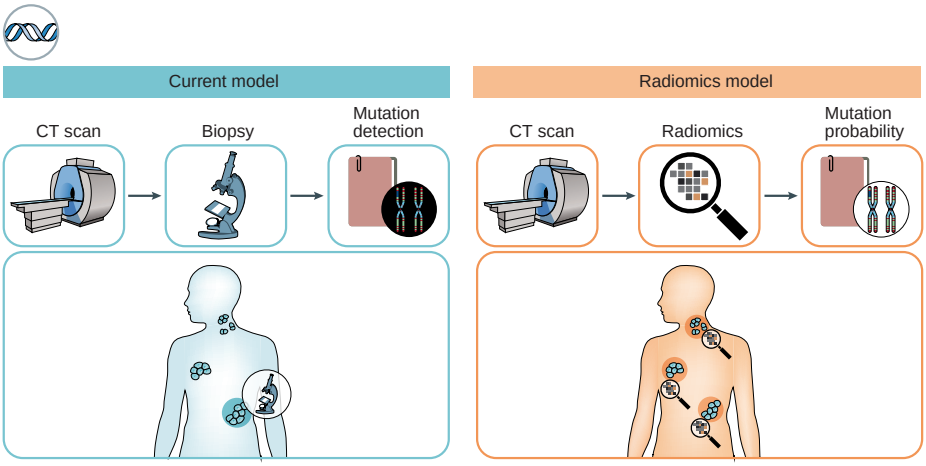


Figure 4 – Radiogenomics analysis may reveal the relationship between imaging phenotypes and gene expression patterns which include expressions of individual genes as well as measures that summarize expressions of specific gene subsets.

THE WAY FORWARD FOR RADIOMICS

Virtual biopsy

In patients with cancer, different parts of the tumours have distinct molecular characteristics and such differences in the tumours change over time. It is impossible to biopsy every part of every tumour at multiple time points; thus, patients tumours are not characterized optimally using biopsies (**Figure 4**).

Delta radiomics

Published work mainly focused on imaging data acquired at a single time point, mostly imaging before the start of treatment. Delta-radiomics introduces a time component and comprises extraction of quantitative features from image sets acquired over the course of treatment [102-104], which provides information on the evolution of feature values [105] (**Figure 5**). Delta-radiomics promises to improve diagnosis, prognosis, prediction, monitoring, image-based intervention or assessment of therapeutic response.

INFRASTRUCTURE FOR RADIOMICS

Radiomics demonstrates huge potential to deepen knowledge and broaden the horizons of imaging toward greater precision and extraction of *in-vivo* biological information. To fully harness the potential of radiomics, the research and clinical communities must embrace an interdisciplinary shared vision of personalized precision medicine. Extracted radiomic features must be stored in searchable databases in order to realize the unprecedented potential for RLHC that routine standard-of-care imaging represents. Hence, RLHC networks can dynamically capture multimodal data and share knowledge across departmental and institutional boundaries, in order to accumulate sufficient datasets for significant statistical power in model development and validation.

Big data

Idealized RLHC necessitates the 4Vs of 'big data'; volume, variety, velocity, and veracity of data. The volume of data is important: first, to gain greater knowledge because the quality of the knowledge correlates with the number of patients on whom that knowledge is founded; second, to enable more variables in the model development phase; third, to gain knowledge regarding rare patient groups. The variety of data, both in terms of treatment and of patient characteristics, is critical for deciding which treatment is optimal for an individual patient. The velocity of data is important to guarantee

that knowledge is gathered as swiftly and perpetually as possible, while the veracity of data is critical to the amount of confidence that can be ascribed to the knowledge gained.

Data sharing

Procuring data of sufficient quality with regard to the 4Vs is central to RLHC. There is a pressing need in both the research and clinical communities to embrace knowledge and data-sharing technology [106], which transcends institutional and national boundaries [107]. The following established obstacles to data sharing [108] are apparent in the medical domain: human resources or insufficient time; cultural and language difficulties, data recording methods; the political and academic value of data; hazards to reputation; legal and privacy considerations, to name a few. These are all significant issues to address and are not easy to overcome.

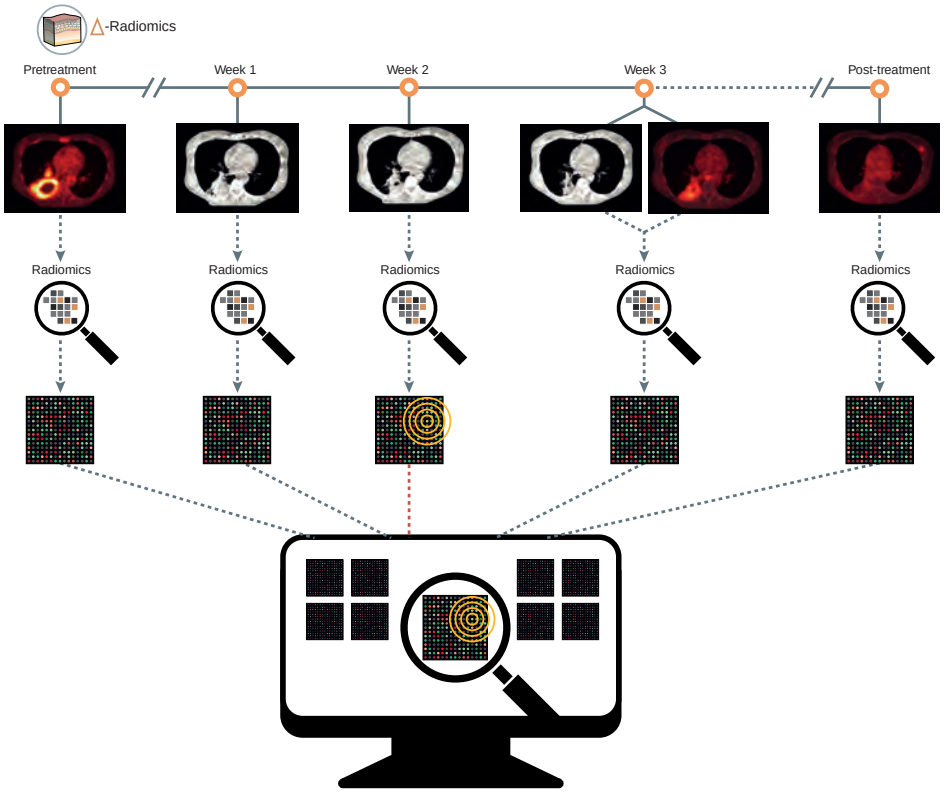


Figure 5 – Schematic overview of a clinical decision support system graphical user interface illustrating the concept of Δ-radiomics, i.e., a clinician user requests the radiomic analysis of a patient based upon combined longitudinal PET/CT images enabling potentially: improved diagnosis, early response prediction, improved clinical decision making, and consequently a better prognosis.

One initiative to accomplish this goal is CancerLinQ [109], the ASCO data centralization approach. Another initiative is worldCAT that consists of a novel data-federated approach that successfully links radiotherapy institutes in the Netherlands, Germany, Belgium, the UK (**Figure 6**), Italy, Australia, China, India, and the USA (Supplementary video [110]). In addition, universal streamlined solutions through advanced information communication technologies have been central to the realization of this endeavor, readily facilitating synchronized RLHC in each centre without inclusion of sensitive data, which overcomes the classic barriers to data sharing. Other important links include the cancer imaging archive (TCIA) [111], The Quantitative Imaging Network (QIN) [112], the Quantitative Imaging Biomarkers Alliance (QIBA) [113], and quantitative imaging in cancer by connecting cellular processes with therapy (QuIC-ConCePT) [114].

Ontologies for learning

For RLHC to succeed, the creation of data with semantic interoperability, also known as 'machine-readable' data [115] is needed, in which local terms are harmonized from concepts of well-defined ontologies (such as the NCI Thesaurus or ICD-10). Exploiting this technique, the ontology terms serve as a common reference for the data at each institutional site, permitting a unified process for information retrieval enabled by a semantic gateway to the data. A benefit of this approach is that it promotes standardization with respect to data management (such as disease-specific 'umbrella' protocols: NCT01855191) [116, 117].

The role of radiomics in the future

Picture archiving and radiomics knowledge systems (PARKS) of the future will identify, segment, and extract features from regions of interest. If previous images associated with the same patient are accessible, the earlier identified regions of interest will be automatically identified by the PARKS software. The PARKS will automatically extract quantitative image features that are uploaded to a shared database and compared with previous images to enable more powerful 'learning' to enhance CDSS for diagnosis, prognosis, and treatment, resulting in improved personalization and precision medicine (**Figure 7**). Such capabilities are on the technological, scientific, and clinical horizons, as most current picture archiving and communication systems have the capability to co-register current images with previous images and perform user-interactive segmentation. For the immediate future, the field of radiomics will be focused on the creation of suitable infrastructures for powerful RLHC networks that will facilitate the development and validation of models.

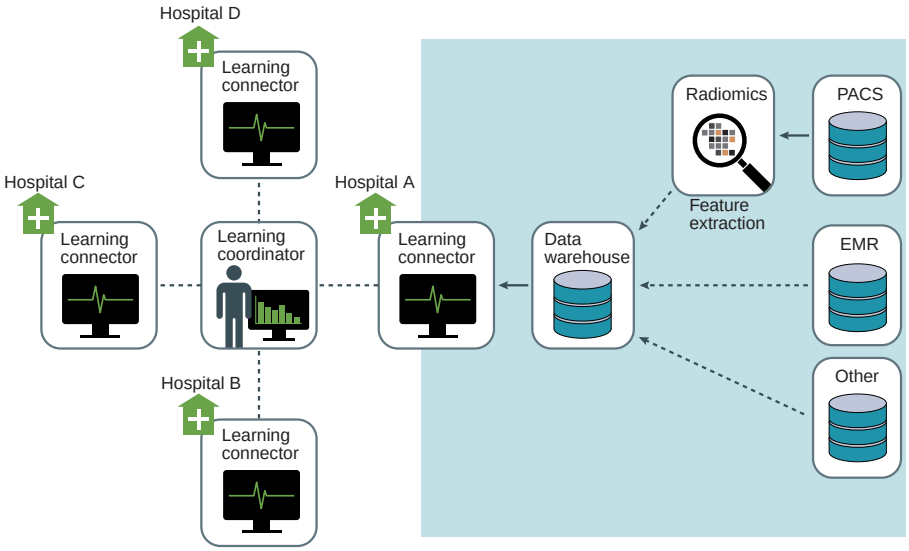


Figure 6 – Schematic diagram of the CAT system. Multiple centers are linked via their Learning Connectors. The connector is the interface where machine learning algorithms, which are sent from the learning coordinator, learn models from local data but privacy-sensitive information never leaves the institute. Partner sites exist in the Netherlands, Germany, Belgium, Italy, Denmark, Australia, China, India, Ireland, UK, and the USA. The system is built from a combination of open source information communication technologies and can deliver data locally via SQL query, or to the wider CAT network via a SPARQL endpoint.

CONCLUSIONS

Our vision for radiomics is expansive and bold. In the reasonably near future, we envision that CDSS that apply knowledge leveraged from radiomic features mined from global RLHC networks populated by (standard-of-care) imaging will enable increased personalization and precision within medicine. For this vision to be actualized within the routine clinical setting, clinicians and medical physicists must be incentivized to participate in the process; standardization is crucial to this endeavor, principally in high-quality data acquisition (clinical, treatment, imaging, genetic, etc.). Standardization obliges coherent clinical guidelines with agreed standards for image acquisition and analysis, as well as data-sharing techniques that exploit matching ontologies. Continuous re-evaluation and demonstrating the clinical utility of a CDSS is as significant as standardizing the development and validation of the design of clinical trials. These crucial steps are the foundations of a successful CDSS. Simultaneous and synergistic advances in RLHC and radiomics will empower the next major breakthroughs in personalization and precision medicine.

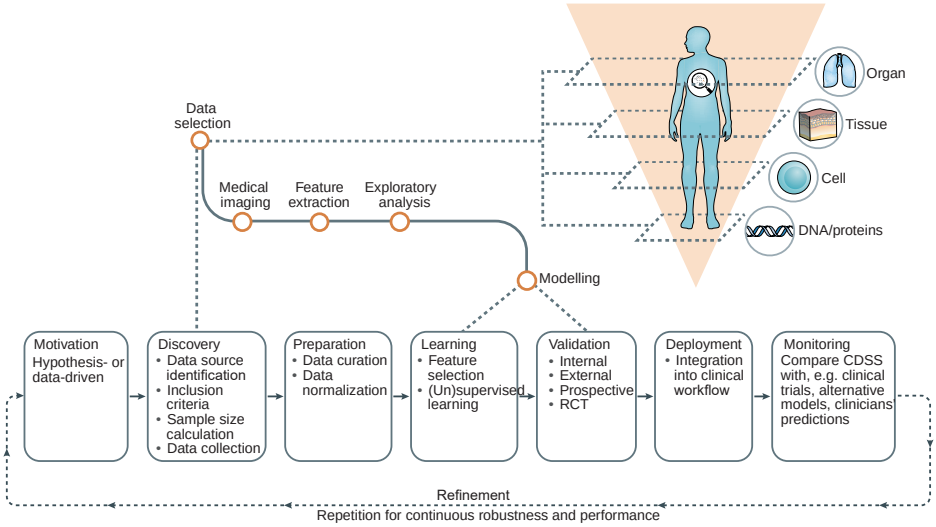


Figure 7 – Overview of the methodological processes for RLHC and how the radiomics workflow fits into the development of a CDSS: Data selection, discovery, collection and preparation, model(s) development/validation and implementation, assessment of clinical utility and ultimately refinement through continuous repetition of the process (quality control and assurance protocols are requisite throughout the process).

ACKNOWLEDGMENTS

The authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015, no. 694812) and the QuIC-ConCePT project, which is partly funded by EFPI A companies and the Innovative Medicine Initiative Joint Undertaking (IMI JU) under Grant Agreement no. 115151. This research is also supported by the Dutch Technology Foundation STW (grant no. 10696 duCAT & P14-19 Radiomics STRaTegy), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs. Authors also acknowledge financial support from the National Institute of Health (NIH-USA U01 CA 143062-01, Radiomics of NSCLC), EU 7th framework program (EURECA, ARTFORCE – no. 257144, REQUITE – no. 601826), SME Phase 2 (EU proposal 673780 – RAIL), the European Program H2020 (BD2Decide – PHC30-689715, ImmunoSABR – no. 733008, PREDICT - ITN no. 766276), Kankeronderzoekfonds Limburg from the Health Foundation Limburg and the Dutch Cancer Society (KWF UM 2011-5020, KWF UM 2009-4454, KWF MAC 2013-6425, KWF MAC 2013-6089) and Alpe d'HuZes-KWF (DESIGN), Center for Translational Molecular Medicine (TraIT), EUROSTARS (SeDI, CloudAtlas, and DART), Interreg V-A Euregio Meuse-Rhine ("Euradiomics") and Varian Medical Systems (VATE and ROO).

Table 1 – The radiomics quality score: RQS

	Criteria	Points
1	Image protocol quality - well-documented image protocols (e.g., contrast, slice thickness, energy, etc.) and/or usage of public image protocols allow reproducibility/replicability.	+ 1 (if protocols are well-documented) + 1 (if public protocol is used)
2	Multiple segmentations - possible actions are: segmentation by different physicians/algorithms/software, perturbing segmentations by (random) noise, segmentation at different breathing cycles. Analyze feature robustness to segmentation variabilities.	+ 1
3	Phantom study on all scanners - detect inter-scanner differences and vendor-dependent features. Analyze feature robustness to these sources of variability.	+ 1
4	Imaging at multiple time points - collect individuals' images at additional time points. Analyze feature robustness to temporal variabilities (e.g., organ movement, organ expansion/shrinkage).	+ 1
5	Feature reduction or adjustment for multiple testing - decreases the risk of overfitting. Overfitting is inevitable if the number of features exceeds the number of samples. Consider feature robustness when selecting features.	- 3 (if neither measure is implemented) + 3 (if either measure is implemented)
6	Multivariable analysis with non-radiomics features (e.g., EGFR mutation) - is expected to provide a more holistic model. Permits correlating/inferencing between radiomics and non-radiomics features.	+ 1
7	Detect and discuss biological correlates - demonstration of phenotypic differences (possibly associated with underlying gene-protein expression patterns) deepens understanding of radiomics and biology.	+ 1
8	Cut-off analyses - determine risk groups by either the median, a previously published cut-off or report a continuous risk variable. Reduces the risk of reporting overly optimistic results.	+ 1
9	Discrimination statistics - report discrimination statistics (e.g., C-statistic, ROC curve, AUC) and their statistical significance (e.g., p-values, confidence intervals). One can also apply resampling method (e.g., bootstrapping, cross-validation).	+ 1 (if a discrimination statistic and its statistical significance are reported) + 1 (if also an resampling method technique is applied)
10	Calibration statistics - report calibration statistics (e.g., Calibration-in-the-large/slope, calibration plots) and their statistical significance (e.g., p-values, confidence intervals). One can also apply resampling method (e.g., bootstrapping, cross-validation).	+ 1 (if a calibration statistic and its statistical significance are reported) + 1 (if also an resampling method technique is applied)
11	Prospective study registered in a trial database - provides the highest level of evidence supporting the clinical validity and usefulness of the radiomics biomarker.	+ 7 (for prospective validation of a radiomics signature in an appropriate trial)

12	Validation - the validation is performed without retraining and without adaptation of the cut-off value, provides crucial information with regard to credible clinical performance.	- 5 (if validation is missing) + 2 (if validation is based on a dataset from the same institute) + 3 (if validation is based on a dataset from another institute) + 4 (if validation is based on two datasets from two distinct institutes) + 4 (if the study validates a previously published signature) + 5 (if validation is based on three or more datasets from distinct institutes) *Datasets should be of comparable size and should have at least 10 events per model feature.
13	Comparison to 'gold standard' - assess the extent to which the model agrees with/is superior to the current 'gold standard' method (e.g., TNM-staging for survival prediction). This comparison shows the added value of radiomics.	+ 2
14	Potential clinical utility - report on the current and potential application of the model in a clinical setting (e.g., decision curve analysis).	+ 2
15	Cost-effectiveness analysis - report on the cost-effectiveness of the clinical application (e.g., QALYs generated).	+ 1
16	Open science and data - make code and data publicly available. Open science facilitates knowledge transfer and reproducibility of the study.	+ 1 (if scans are open source) + 1 (if region of interest segmentations are open source) + 1 (if code is open source) + 1 (if radiomics features are calculated on a set of representative ROIs and the calculated features + representative ROIs are open source)
	Total points (36 = 100%)	

Table 2 – Radiomics in practice

Utility	Modality	Features	Cancer	#Pts	Result	Conclusion	Ref
Tumour prognosis	CT	Intensity, shape, texture, and wavelet	Lung and head & neck	1,019	Lung: (C-index = 0.65, p-value = 2.9×0^{-09}, Wilcoxon test), and a high performance in H&N1 (C-index = 0.69, p-value = 8.0×10^{-07}, Wilcoxon test) and H&N2 (C-index = 0.69, p-value = 3.5×10^{-06}, Wilcoxon test).	Could predict survival in two entirely independent external cohorts of patients, outperforming the current gold standard of tumor-node-metastasis status (radiation or concurrent chemoradiation).	[1]
Tumour prognosis	PET	Texture and shape	Esophageal	217	A clinical prediction model (C-index = 0.67) was improved by adding radiomic features (C-index = 0.77). However, at a decision threshold of ≥0.9 there was no clear incremental value.	Demonstrated that a radiomic PET signature provided statistical incremental value for predicting pathological complete response after preoperative chemoradiotherapy.	[118]
Tumor prognosis	CT		Colorectal	326	Training: showed good discrimination (C-index = 0.74) and calibration. Validation: showed good discrimination (C-index = 0.78) and good calibration.	Decision curve analysis demonstrated that a final nomogram consisting of the radiomic portal venous-phase CT signature, CT-reported lymph node status, and carcinoembryonic antigen level was clinically useful.	[119]

Distant metastasis	CT	Texture, Laplacian of Gaussian and wavelet filters	Lung	182	Could predict distant metastasis in an independent validation dataset (C-index = 0.61, p-value = 1.79×10^{-17}). Adding this radiomic-signature to a clinical model resulted in a significant improvement (p-value = 1.56×10^{-11}).	Provided superior information than clinical data capturing detailed information of the tumor phenotype and can be used as a prognostic biomarker for distant metastasis.	[120]
Distant metastasis	CT	Wavelet filters	Lung	113	Significantly prognostic for distant metastasis (C-index = 0.67, q-value < 0.1), while none of the conventional and clinical parameters were. Three conventional and four radiomic features were prognostic for overall survival.	Demonstrates that radiomic features have potential to be prognostic for some outcomes that conventional imaging metrics cannot predict in stereotactic body radiation therapy patients	[121]
Efficacy	CT	Intensity and texture	Esophageal	106	Significant change in radiomics feature values was observed with increasing radiation dose (pre and post radiotherapy scans). AUC = 0.75 using multiple features in a classifier	Demonstrated the ability to individualized measurement of patient lung tissue reaction to radiotherapy and assess radiation pneumonitis development.	[122]

Staging	CT	Intensity, texture, Laplacian of Gaussian filters	Colorectal	494	Training: AUC = 0.79, p-value = <0.0001. Validation: AUC = 0.71, p-value = <0.0001. The radiomics signature was an independent predictor for staging.	Demonstrated the ability to discriminate between stage I-II from III-IV, which may serve as a complementary tool for the preoperative tumor staging.	[123]
Screening	CT	Intensity and texture	Lung	196	AUC = 0.83, p-value = <0.05. Radiomics performance was commensurate with the McWilliams risk assessment model.	Demonstrated that radiomics at baseline can be used to assess risk for development of cancer.	[124]
Survival	MR	Volumetric	Brain	141	C-index=0.60, p-value=$4x10^{-4}$. Volumetric features were significantly associated with diverse sets of biological processes, FDR<0.05.	Demonstrated the ability to derive the biological state of a glioblastoma tumor that can be used to develop personalized treatment strategies.	[125]
Survival	CT	Texture	Lung	282	C-index=0.72, improved accuracy of calibration and the classification of survival outcomes (net reclassification improvement: 0.182, p-value = .02).	Decision curve analysis demonstrated that in terms of clinical usefulness, the radiomics nomogram outperformed the traditional staging system and the clinical-pathologic nomogram.	[126]

Tumor prognosis	CT	Intensity, shape, texture, and wavelet	Oropharyngeal	542	C-index=0.63, p-value=2.72 $x10^{-9}$. Kaplan-Meier survival curves were significantly different (p-value < 0.05) between high and low radiomic signature model predictions for all cohorts.	Demonstrated external validation of the signature, the signature had significant prognostic power irrespective of the presence or not of CT artifacts.	[127]
Overall survival	PET	Shape, intensity, and texture	Pancreatic	139	C-index=0.66, significantly better than competing prognostic indices (0.48-0.64, Wilcoxon rank sum test p-value=1 $x10^{-6}$).	Demonstrated external validation of the signature, If validated in large, prospective cohorts, the signature might be used to identify patients for individualized risk-adaptive therapy.	[128]
Recurrence	MR	Shape, intensity, and texture	Breast	89	AUC = 0.88, 0.76, and 0.68 for MammaPrint, Oncotype DX, and PAM50 risk of relapse based on subtype respectively, all statistically significant, , p-value = ≤.05.	Demonstrates that breast MR imaging radiomics shows promise for image-based phenotyping in assessing the risk of breast cancer recurrence.	[129]

REFERENCES

[1] Aerts H, Rios-Velazquez E, Leijenaar R, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Comms. 2014;5.

[2] Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nature reviews. Clinical oncology 2011;8: 184-187.

[3] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48: 441-446.

[4] Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magnetic resonance imaging 2012;30: 1234-1248.

[5] Haase AT, Henry K, Zupancic M, Sedgewick G, Faust RA, Melroe H, et al. Quantitative Image Analysis of HIV-1 Infection in Lymphoid Tissue. Science 1996;274: 985-989.

[6] Lambin P, van Stiphout R, Starmans M, Rios-Velazquez E, Nalbantov G, Aerts H. Predicting outcomes in radiation oncology–multifactorial decision support systems. Nat. Rev. Clin. Oncol. 2013;10.

[7] Medicine: Computers by the Bedside. Nature 1969;224: 636-637.

[8] Schoolman H, Bernstein L. Computer use in diagnosis, prognosis, and therapy. Science 1978;200: 926-931.

[9] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2015: 151169.

[10] Roelofs E, Dekker A, Meldolesi E, van Stiphout RG, Valentini V, Lambin P. International data-sharing for radiotherapy research: an open-source based infrastructure for multicentric clinical data mining. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology 2014;110: 370-374.

[11] Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology 2013;108: 174-179.

[12] Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific reports 2016;6: 26094.

[13] Nead KT, Gaskin G, Chester C, Swisher-McClure S, Dudley JT, Leeper NJ, et al. Androgen Deprivation Therapy and Future Alzheimer's Disease Risk. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2016;34: 566-571.

[14] Gatenby RA, Grove O, Gillies RJ. Quantitative Imaging in Cancer Evolution and Ecology. Radiology 2013;269: 8-14.

[15] Aerts HL. The potential of radiomic-based phenotyping in precision medicine: A review. JAMA Oncology 2016.

[16] Lambin P, Zindler J, Vanneste B, Van De Voorde L, Eekers D, Compter I, et al. Decision Support Systems for Personalized and Participative Radiation Oncology. Advanced Drug Delivery Reviews 2016.

[17] Vickers A. Prediction models: revolutionary in principle, but do they do more good than harm? J. Clin. Oncol. 2011;29: 2951–2952.

[18] Yip SS, Aerts HJ. Applications and limitations of radiomics. Phys Med Biol 2016;61: R150-166.

[19] Polan DF, Brady SL, Kaufman RA. Tissue segmentation of computed tomography images using a Random Forest algorithm: a feasibility study. Phys Med Biol 2016;61: 6553-6569.

[20] Balagurunathan Y, Gu Y, Wang H, Kumar V, Grove O, Hawkins S, et al. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. Translational Oncology 2014;7: 72-87.

[21] Grootjans W, Tixier F, van der Vos CS, Vriens D, Le Rest CC, Bussink J, et al. The impact of optimal respiratory gating and image noise on evaluation of intra-tumor heterogeneity in 18F-FDG positron emission tomography imaging of lung cancer. Journal of nuclear medicine : official publication, Society of Nuclear Medicine 2016.

[22] Larue RTHM, Defraene G, Ruysscher DD, Lambin P, Elmpt Wv. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. The British Journal of Radiology 2017;90: 20160665.
[23] Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest Radiol 2015;50: 757-765.
[24] Balagurunathan Y, Kumar V, Gu Y, Kim J, Wang H, Liu Y, et al. Test–Retest Reproducibility Analysis of Lung CT Image Features. Journal of Digital Imaging 2014;27: 805-823.
[25] Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, et al. Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non–Small Cell Lung Cancer. Radiology 2009;252: 263-272.
[26] Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Scientific reports 2016;6: 23428.
[27] Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present... any future? European journal of nuclear medicine and molecular imaging 2016.
[28] Fang YH, Lin CY, Shih MJ, Wang HM, Ho TY, Liao CT, et al. Development and evaluation of an open-source software package "CGITA" for quantifying tumor heterogeneity with molecular images. Biomed Res Int 2014;2014: 248505.
[29] Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Medical physics 2015;42: 1341-1353.
[30] Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine Learning methods for Quantitative Radiomic Biomarkers. Scientific reports 2015;5: 13087.
[31] https://github.com/ (accessed on: 2017 May 18th)
[32] Collins G, Reitsma J, Altman D, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. Ann. Intern. Med. 2015;162: 55-63.
[33] Lemeshow S, Hosmer DW, Jr. A review of goodness of fit statistics for use in the development of logistic regression models. American journal of epidemiology 1982;115: 92-106.
[34] Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68: 279-289.
[35] Steyerberg E, Vickers A, Cook N, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21: 128-138.
[36] Leek JT, Peng RD. Statistics: P values are just the tip of the iceberg. Nature 2015;520: 612.
[37] Drummond C. Replicability is not Reproducibility: Nor is it Good Science. In: Evaluation Methods for Machine Learning, 2009.
[38] Peng RD. Reproducible Research in Computational Science. Science 2011;334: 1226-1227.
[39] Peng RD, Dominici F, Zeger SL. Reproducible Epidemiologic Research. American journal of epidemiology 2006;163: 783-789.
[40] https://www.cancerdata.org/resource/doi%3A10.17195/candat.2016.08.1 (accessed on: 2017 May 18th)
[41] http://www.radiomics.world/ (accessed on: 2017 May 18th)
[42] Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. BMC Medicine 2012;10: 1-39.
[43] Pepe MS, Feng Z. Improving Biomarker Identification with Better Designs and Reporting. Clinical chemistry 2011;57: 1093-1095.
[44] Poste G. Biospecimens, biomarkers, and burgeoning data: the imperative for more rigorous research standards. Trends in molecular medicine 2012;18: 717-722.
[45] Rosenstein BS, West CM, Bentzen SM, Alsner J, Andreassen CN, Azria D, et al. Radiogenomics: radiobiology enters the era of big data and team science. Int J Radiat Oncol Biol Phys 2014;89: 709-713.
[46] Rutman AM, Kuo MD. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. European journal of radiology 2009;70: 232-241.

[47] Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proceedings of the National Academy of Sciences of the United States of America 2005;102: 3738-3743.
[48] Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, et al. Gene expression patterns in human liver cancers. Molecular biology of the cell 2002;13: 1929-1939.
[49] Chung CH, Bernard PS, Perou CM. Molecular portraits and the family tree of cancer. Nature genetics 2002;32 Suppl: 533-540.
[50] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. The New England journal of medicine 2004;351: 2817-2826.
[51] Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2006;24: 3726-3734.
[52] Segal E, Friedman N, Kaminski N, Regev A, Koller D. From signatures to models: understanding cancer using microarrays. Nature genetics 2005;37 Suppl: S38-45.
[53] Diehn M, Nardini C, Wang DS, McGovern S, Jayaraman M, Liang Y, et al. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. Proceedings of the National Academy of Sciences of the United States of America 2008;105: 5213-5218.
[54] Gevaert O, Xu J, Hoang CD, Leung AN, Xu Y, Quon A, et al. Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data--methods and preliminary results. Radiology 2012;264: 387-396.
[55] Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nature biotechnology 2007;25: 675-680.
[56] Gao X, Chu C, Li Y, Lu P, Wang W, Liu W, et al. The method and efficacy of support vector machine classifiers based on texture features and multi-resolution histogram from 18F-FDG PET-CT images for the evaluation of mediastinal lymph nodes in patients with lung cancer. European journal of radiology 2015;84: 312-317.
[57] Harry VN, Semple SI, Parkin DE, Gilbert FJ. Use of new imaging techniques to predict tumour response to therapy. The Lancet. Oncology 2010;11: 92-102.
[58] O'Connor JP, Jackson A, Asselin MC, Buckley DL, Parker GJ, Jayson GC. Quantitative imaging biomarkers in the clinical development of targeted therapeutics: current and future perspectives. The Lancet. Oncology 2008;9: 766-776.
[59] Panth KM, Leijenaar RT, Carvalho S, Lieuwes NG, Yaromina A, Dubois L, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology 2015.
[60] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA: a cancer journal for clinicians 2011;61: 69-90.
[61] Wang J, Kato F, Oyama-Manabe N, Li R, Cui Y, Tha KK, et al. Identifying Triple-Negative Breast Cancer Using Background Parenchymal Enhancement Heterogeneity on Dynamic Contrast-Enhanced MRI: A Pilot Radiomics Study. PloS one 2015;10: e0143308.
[62] Abernethy A, Etheredge L, Ganz P, Wallace P, German R, Neti C, et al. Rapid-learning system for cancer care. J Clin Oncol. 2010;28: 4268-4274.
[63] Lambin P, Zindler J, Vanneste B, van de Voorde L, Jacobs M, Eekers D, et al. Modern clinical research: How rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine. Acta Oncol 2015;54: 1289-1300.
[64] Dekker A, Vinod S, Holloway L, Oberije C, George A, Goozee G, et al. Rapid learning in practice: A lung cancer survival decision support system in routine patient care data. Rad. Onc. 2014;113: 7.
[65] Ginsburg G, Staples J, Abernethy A. Academic medical centers: ripe for rapid-learning personalized health care. Sci. Transl. Med. 2011;Sc3: 101-127.
[66] Lambin P, Roelofs E, Reymen B, Velazquez E, Buijsen J, Zegers C, et al. Rapid Learning health care in oncology - An approach towards decision support systems enabling customised radiotherapy. Radiother. Oncol. 2013;109: 159-164.

[67] Buettner R, Wolf J, Thomas RK. Lessons learned from lung cancer genomics: the emerging concept of individualized diagnostics and treatment. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2013;31: 1858-1865.

[68] Colen R, Foster I, Gatenby R, Giger ME, Gillies R, Gutman D, et al. NCI Workshop Report: Clinical and Computational Requirements for Correlating Imaging Phenotypes with Genomics Signatures. Transl Oncol 2014;7: 556-569.

[69] Rizzo S, Petrella F, Buscarino V, De Maria F, Raimondi S, Barberis M, et al. CT Radiogenomic Characterization of EGFR, K-RAS, and ALK Mutations in Non-Small Cell Lung Cancer. European radiology 2015.

[70] Taguchi F, Solomon B, Gregorc V, Roder H, Gray R, Kasahara K, et al. Mass spectrometry to classify non-small-cell lung cancer patients for clinical outcome after treatment with epidermal growth factor receptor tyrosine kinase inhibitors: a multicohort cross-institutional study. Journal of the National Cancer Institute 2007;99: 838-846.

[71] Yaromina A, Krause M, Baumann M. Individualization of cancer treatment from radiotherapy perspective. Molecular oncology 2012;6: 211-221.

[72] Dancey JE, Dobbin KK, Groshen S, Jessup JM, Hruszkewycz AH, Koehler M, et al. Guidelines for the development and incorporation of biomarker studies in early clinical trials of novel agents. Clinical cancer research : an official journal of the American Association for Cancer Research 2010;16: 1745-1755.

[73] Krause M, Yaromina A, Eicheler W, Koch U, Baumann M. Cancer stem cells: targets and potential biomarkers for radiotherapy. Clinical cancer research : an official journal of the American Association for Cancer Research 2011;17: 7224-7229.

[74] Lindegaard JC, Overgaard J, Bentzen SM, Pedersen D. Is there a radiobiologic basis for improving the treatment of advanced stage cervical cancer? Journal of the National Cancer Institute. Monographs 1996: 105-112.

[75] Yaromina A, Krause M, Thames H, Rosner A, Krause M, Hessel F, et al. Pre-treatment number of clonogenic cells and their radiosensitivity are major determinants of local tumour control after fractionated irradiation. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology 2007;83: 304-310.

[76] Lambin P, Petit SF, Aerts HJ, van Elmpt WJ, Oberije CJ, Starmans MH, et al. The ESTRO Breur Lecture 2009. From population to voxel-based radiotherapy: exploiting intra-tumour and intra-organ heterogeneity for advanced treatment of non-small cell lung cancer. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology 2010;96: 145-152.

[77] Prokopiou S, Moros EG, Poleszczuk J, Caudell J, Torres-Roca JF, Latifi K, et al. A proliferation saturation index to predict radiation response and personalize radiotherapy fractionation. Radiation Oncology 2015;10: 1-8.

[78] Yin Q, Hung SC, Wang L, Lin W, Fielding JR, Rathmell WK, et al. Associations between Tumor Vascularity, Vascular Endothelial Growth Factor Expression and PET/MRI Radiomic Signatures in Primary Clear-Cell-Renal-Cell-Carcinoma: Proof-of-Concept Study. Scientific reports 2017;7: 43356.

[79] Menegakis A, De Colle C, Yaromina A, Hennenlotter J, Stenzl A, Scharpf M, et al. Residual gammaH2AX foci after ex vivo irradiation of patient samples with known tumour-type specific differences in radio-responsiveness. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology 2015.

[80] Menegakis A, von Neubeck C, Yaromina A, Thames H, Hering S, Hennenlotter J, et al. gammaH2AX assay in ex vivo irradiated tumour specimens: A novel method to determine tumour radiation sensitivity in patient-derived material. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology 2015.

[81] Slonina D, Gasinska A. Intrinsic radiosensitivity of healthy donors and cancer patients as determined by the lymphocyte micronucleus assay. International journal of radiation biology 1997;72: 693-701.

[82] Fertil B, Malaise EP. Intrinsic radiosensitivity of human cell lines is correlated with radioresponsiveness of human tumors: analysis of 101 published survival curves. International journal of radiation oncology, biology, physics 1985;11: 1699-1707.

[83] Menegakis A, Yaromina A, Eicheler W, Dorfler A, Beuthien-Baumann B, Thames HD, et al. Prediction of clonogenic cell survival curves based on the number of residual DNA double strand breaks measured by gammaH2AX staining. International journal of radiation biology 2009;85: 1032-1041.
[84] Bjork-Eriksson T, West C, Karlsson E, Mercke C. Tumor radiosensitivity (SF2) is a prognostic factor for local control in head and neck cancers. International journal of radiation oncology, biology, physics 2000;46: 13-19.
[85] Chitnis MM, Lodhia KA, Aleksic T, Gao S, Protheroe AS, Macaulay VM. IGF-1R inhibition enhances radiosensitivity and delays double-strand break repair by both non-homologous end-joining and homologous recombination. Oncogene 2014;33: 5262-5273.
[86] Du S, Bouquet S, Lo CH, Pellicciotta I, Bolourchi S, Parry R, et al. Attenuation of the DNA damage response by transforming growth factor-beta inhibitors enhances radiation sensitivity of non-small-cell lung cancer cells in vitro and in vivo. International journal of radiation oncology, biology, physics 2015;91: 91-99.
[87] Kahn J, Hayman TJ, Jamal M, Rath BH, Kramp T, Camphausen K, et al. The mTORC1/mTORC2 inhibitor AZD2014 enhances the radiosensitivity of glioblastoma stem-like cells. Neuro-oncology 2014;16: 29-37.
[88] West CM, Davidson SE, Roberts SA, Hunter RD. The independence of intrinsic radiosensitivity as a prognostic factor for patient response to radiotherapy of carcinoma of the cervix. British journal of cancer 1997;76: 1184-1190.
[89] Cheng Q, Roelofs E, Ramaekers BLT, Eekers D, van Soest J, Lustberg T, et al. Development and evaluation of an online three-level proton vs photon decision support prototype for head and neck cancer – Comparison of dose, toxicity and cost-effectiveness. Radiotherapy and Oncology 2016;118: 281-285.
[90] Okada H, Weller M, Huang R, Finocchiaro G, Gilbert MR, Wick W, et al. Immunotherapy response assessment in neuro-oncology: a report of the RANO working group. The Lancet. Oncology 2015;16: e534-542.
[91] Tang C, Amer A, Hobbs B, Li X, Behrens C, Para Cuentas E, et al. Pathology-Based Non-Small Cell Lung Cancer Radiomics Signature Describing the Local Tumor Immune Environment: Discovery and Validation. International Journal of Radiation Oncology • Biology • Physics 2016;96: S42-S43.
[92] Formenti SC, Demaria S. Combining radiotherapy and cancer immunotherapy: a paradigm shift. Journal of the National Cancer Institute 2013;105: 256-265.
[93] Coulie PG, Van den Eynde BJ, van der Bruggen P, Boon T. Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. Nature reviews. Cancer 2014;14: 135-146.
[94] Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science 2015;348: 69-74.
[95] Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell 2015;160: 48-61.
[96] Mellman I, Steinman RM. Dendritic cells: specialized and regulated antigen processing machines. Cell 2001;106: 255-258.
[97] Demaria S, Golden EB, Formenti SC. Role of Local Radiation Therapy in Cancer Immunotherapy. JAMA Oncol 2015.
[98] Golden EB, Chhabra A, Chachoua A, Adams S, Donach M, Fenton-Kerimian M, et al. Local radiotherapy and granulocyte-macrophage colony-stimulating factor to generate abscopal responses in patients with metastatic solid tumours: a proof-of-principle trial. The Lancet. Oncology 2015;16: 795-803.
[99] Garon EB, Rizvi NA, Hui R, Leighl N, Balmanoukian AS, Eder JP, et al. Pembrolizumab for the treatment of non-small-cell lung cancer. The New England journal of medicine 2015;372: 2018-2028.
[100] Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 2015;348: 124-128.
[101] Sanghera S, Barton P, Bhattacharya S, Horne AW, Roberts TE. Pharmaceutical treatments to prevent recurrence of endometriosis following surgery: a model-based economic evaluation. BMJ Open 2016;6: e010580.
[102] Carvalho S, Leijenaar RTH, Troost EGC, van Elmpt W, Muratet JP, Denis F, et al. Early variation of FDG-PET radiomics features in NSCLC is related to overall survival - the “delta radiomics” concept. Radiotherapy and Oncology 2016;118: S20-S21.

[103]Fave X, Mackin D, Yang J, Zhang J, Fried D, Balter P, et al. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? Medical physics 2015;42: 6784-6797.
[104]Leijenaar RTH, Nalbantov G, Carvalho S, van Elmpt WJC, Troost EGC, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Scientific reports 2015;5: 11075.
[105]Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Scientific reports 2017;7: 588.
[106]Deasy JO, Bentzen SM, Jackson A, Ten Haken RK, Yorke ED, Constine LS, et al. Improving Normal Tissue Complication Probability Models: The Need to Adopt a "Data-Pooling" Culture. International journal of radiation oncology, biology, physics 2010;76: S151-S154.
[107]Skripcak T, Belka C, Bosch W, Brink C, Brunner T, Budach V, et al. Creating a data exchange strategy for radiotherapy research: Towards federated databases and anonymised public datasets. Radiother. Oncol. 2014;113: 303-309.
[108]Budin-Ljosne I, Burton P, Isaeva J, Gaye A, Turner A, Murtagh MJ, et al. DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis. Public health genomics 2015;18: 87-96.
[109]Schilsky RL, Michels DL, Kearbey AH, Yu PP, Hudis CA. Building a rapid learning health care system for oncology: the regulatory framework of CancerLinQ. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2014;32: 2373-2379.
[110]http://youtu.be/ZDJFOxpwqEA (accessed on: 2017 May 18rh)
[111]http://www.cancerimagingarchive.net/ (accessed on: 2017 May 18th)
[112]http://imaging.cancer.gov/programsandresources/specializedinitiatives/qin (accessed on: 2017 May 18th)
[113]https://www.rsna.org/qiba/ (accessed on: 2017 May 18th)
[114]http://www.quic-concept.eu/ (accessed on: 2017 May 18th)
[115]Benedict SH, Hoffman K, Martel MK, Abernethy AP, Asher AL, Capala J, et al. Overview of the American Society for Radiation Oncology–National Institutes of Health–American Association of Physicists in Medicine Workshop 2015: Exploring Opportunities for Radiation Oncology in the Era of Big Data. International Journal of Radiation Oncology • Biology • Physics 2016;95: 873-879.
[116]Meldolesi E, van Soest J, Dinapoli N, Dekker A, Damiani A, Gambacorta M, et al. An umbrella protocol for standardized data collection (SDC) in rectal cancer: a prospective uniform naming and procedure convention to support personalized medicine. Radiother. Oncol. 2014;112: 59-62.
[117]https://www.cancerdata.org/protocols/eurocat-umbrella-protocol-nsclc (accessed on: 2017 May 18th)
[118]van Rossum PS, Fried DV, Zhang L, Hofstetter WL, van Vulpen M, Meijer GJ, et al. The incremental value of subjective and quantitative assessment of 18F-FDG PET for the prediction of pathologic complete response to preoperative chemoradiotherapy in esophageal cancer. Journal of nuclear medicine : official publication, Society of Nuclear Medicine 2016.
[119]Huang YQ, Liang CH, He L, Tian J, Liang CS, Chen X, et al. Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2016;34: 2157-2164.
[120]Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RT, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 2015;114: 345-350.
[121]Huynh E, Coroller TP, Narayan V, Agrawal V, Hou Y, Romano J, et al. CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology 2016.
[122]Cunliffe A, Armato SG, 3rd, Castillo R, Pham N, Guerrero T, Al-Hallaq HA. Lung texture in serial thoracic computed tomography scans: correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development. International journal of radiation oncology, biology, physics 2015;91: 1048-1056.
[123]Liang C, Huang Y, He L, Chen X, Ma Z, Dong D, et al. The development and validation of a CT-based radiomics signature for the preoperative discrimination of stage I-II and stage III-IV colorectal cancer. Oncotarget 2016.

[124]Hawkins S, Wang H, Liu Y, Garcia A, Stringfield O, Krewer H, et al. Predicting Malignant Nodules from Screening CT Scans. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer 2016.

[125]Grossmann P, Gutman DA, Dunn WD, Jr., Holder CA, Aerts HJ. Imaging-genomics reveals driving pathways of MRI derived volumetric tumor phenotype features in Glioblastoma. BMC cancer 2016;16: 611.

[126]Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, et al. Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non—Small Cell Lung Cancer. Radiology 2016;0: 152234.

[127]Leijenaar RT, Carvalho S, Hoebers FJ, Aerts HJ, van Elmpt WJ, Huang SH, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta Oncol 2015;54: 1423-1429.

[128]Cui Y, Song J, Pollom E, Alagappan M, Shirato H, Chang DT, et al. Quantitative Analysis of ¹⁸F-Fluorodeoxyglucose Positron Emission Tomography Identifies Novel Prognostic Imaging Biomarkers in Locally Advanced Pancreatic Cancer Patients Treated With Stereotactic Body Radiation Therapy. International Journal of Radiation Oncology • Biology • Physics 2016;96: 102-109.

[129]Li H, Zhu Y, Burnside ES, Drukker K, Hoadley KA, Fan C, et al. MR Imaging Radiomics Signatures for Predicting the Risk of Breast Cancer Recurrence as Given by Research Versions of MammaPrint, Oncotype DX, and PAM50 Gene Assays. Radiology 2016: 152110.

SUPPLEMENTARY MATERIAL

Digital phantom data

Description of the digital phantom image data

To compare different software implementations for radiomic feature extraction algorithms, we provide CT data of the primary tumor region (i.e. a 5 cm margin around the tumor volume) and the corresponding tumor contours of 4 lung cancer cases, to serve as "real life" digital phantoms. The images have an in plane pixel spacing of 0.977 mm and a slice thickness of 3 mm. The data is provided both in original and pre-processed form. All image and contour data is provided in DICOM format and is publicly available (DOI: 10.17195/candat.2016.08.1). The data was generated in Matlab R2014a (The Mathworks, Natick, MA) using an adapted version of CERR [1]. The DICOM RTSTRUCTs, containing the contour coordinates, may be used to compare different implementations to create 3D binary masks from polygon data.

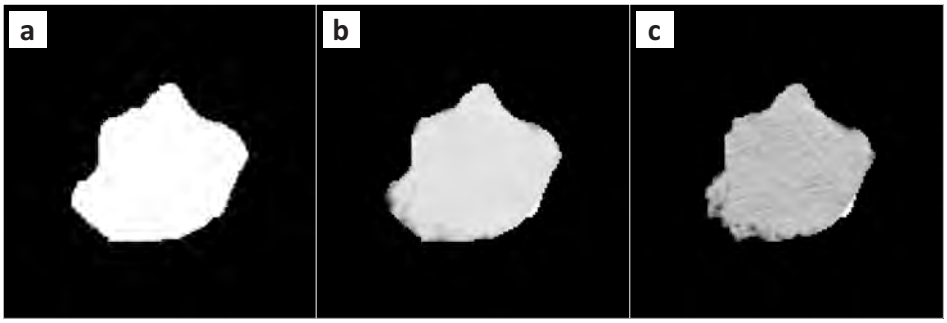


Figure 1 – **(a)** The 3D binary mask I_M. **(b)** The pre-processed image with an offset of + 1000 HU (I_O). **(c)** The pre-processed gray value image with a bin width of 25 HU (I_B).

Image pre-processing

To be able to correctly compare between different implementations of feature extraction algorithms, resulting feature values should not be affected by differences in segmentation and intensity discretization, which are part of image pre-processing. To ensure that the same region of interest (ROI) is used for feature calculations–eliminating dependency of feature values on segmentation–the delineation of the primary tumor (GTV) has been converted to a 3D binary ([0,1]) mask image (I_M; **Figure 1a**) and applied to the original image.

We provide two pre-processed images per digital phantom: I_O (**Figure 1b**), the original image with an offset of +1000 Hounsfield Units (HU), eliminating negative values (air = 0 HU), and I_B (**Figure 1c**) a gray value image discretized into equally spaced bins, with a bin

width of 25 HU. Resampling intensity values into bins with an intensity resolution of $B = 25\ HU$ was performed using:

$$I_B(x) = \left\lceil \frac{I(x)}{B} \right\rceil - \min\left(\left\lceil \frac{I(x)}{B} \right\rceil\right) + 1 \quad (1)$$

Where term $[\min(\lceil I(x)/B \rceil) + 1]$ ensures that the bin count starts at 1. Since differences in intensity discretization affect the resulting feature values [2], the gray values of the pre-processed images should therefore be used as-is. Voxels outside the ROI, which should be ignored for feature calculations, are set to -1000 for both I_O and I_B. The in plane pixel spacing and slice thickness have not been altered.

Machine learning

Machine learning has provided self-driving cars, practical speech recognition, effective web search, a vastly improved understanding of the human genome, and can be leveraged in radiomics analysis. Given below are resources for two of the most effective software packages (caret and scikit-learn) to perform machine learning through the prevalent coding languages R and Python. These resources will enable the reader to gain knowledge implementing appropriate machine learning techniques though practical know-how, leading to quick, powerful, and appropriate application of machine learning techniques to new problems.

- caret
  - http://topepo.github.io/caret/available-models.html
  - https://cran.r-project.org/web/packages/caret/caret.pdf
- scikit-learn
  - http://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf
  - http://scikit-learn.org/stable/user_guide.html

REFERENCES

[1] Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Med Phys 2003;30: 979-985.

[2] Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Scientific reports 2015;5: 11075.

Chapter 10

Extended discussion and future perspectives

Ralph T.H. Leijenaar

To facilitate further development and acceptance of radiomics, this thesis aimed to provide deeper understanding of several fundamental technical and methodological aspects, in particular related to interoperability (Chapter 3-5). This thesis furthermore investigated potential clinical applications of radiomics with a number of proof of concept studies focusing on lung and head and neck cancer (Chapters 6-8).

Radiomics is a relatively young, yet fast growing research field, which holds great promise in precision medicine. Radiomics is extensively discussed in recent literature [1-10]. However, the work presented in Chapter 9 aimed to go one step futher, by not only providing an extensive review and discussion of radiomics with its challenges and opportunities, but also providing a digital phantom and a radiomics quality score to facilitate standardization, interoperability and advancement of the field. The current chapter provides an extended discussion and additional future prospects concerning the work presented in this thesis.

TECHNICAL AND METHODOLOGICAL ASPECTS OF RADIOMICS

Chapter 3 presented an integrated stability analysis of radiomic features derived from PET imaging of non-small cell lung cancer patients. Feature stability was assessed based on test-retest scans of 11 patients, and on inter-observer reproducibility across independent manual tumor delineations of 5 radiation oncologists for a total of 23 patients. In terms of stability ranking, features more stable in test-retest were in general also found to be more robust against inter-observer variability. In Chapter 6, a similar feature ranking approach was applied to select reliable features derived from CT images for the development of a prognostic radiomic signature. Even though such stability analyses are useful to identify robust and reliable features, additional sources of variability make it difficult to generalize results from a single study. In a more recent study, we illustrate that test-retest results are indeed not generalizable and depend on the data they are derived from [11] (not part of this thesis). There is a dependency on single or a combination of factors that can influence feature stability, such as imaging modality and hardware, scan acquisition and reconstruction settings, disease site, test-retest scan time interval, or respiratory motion [12] which should ideally be independently tested for their influence. It is therefore advisable to perform test–retest analyses that are study specific and controlled for these factors. Phantom studies, as discussed in Chapter 9, can serve as an important and useful tool to isolate and independently assess the effect of different possible sources of variability [13, 14].

Nonetheless, test-retest data is not always available. In another recent study, we therefore proposed to use different phases of respiratory correlated 4DCT scans as an alternative for test-retest data, since 4DCT is routinely acquired for treatment planning purposes [15] (not part of this thesis). A strong agreement between feature stability based on 4DCT and test-retest data in lung cancer suggests that 4DCT could serve as

adequate substitute to assess the stability of radiomic features. If no study specific data is available to evaluate feature stability, a purely data driven approach may be considered instead (Chapter 8). An intricate modeling approach could reduce the risk of selecting, meaningless and unreliable features, since the same association with the outcome of interest is unlikely to be consistently found for an unstable/unreliable feature in different data partitions (e.g. when using repeated cross-validation for model selection), or in independent validation datasets. Even though such a data driven approach may provide satisfactory results, it is nevertheless favorable to include appropriate data to assess feature stability whenever available.

In Chapter 4, two conceptually different methods for image intensity discretization were compared for calculating several widely used textural features. Here, it was concluded that that the manner of image intensity discretization has an effect on resulting textural features and, more importantly, has a crucial impact on their interpretation. The latter suggested that intensity discretization using different bin widths might provide complementary information. A more general approach is therefore to consider features calculated with different bin widths (or number of bins) as different features altogether, meaning that the intensity discretization scheme is part of the feature definitions [14]. In Chapter 8, this was exploited by calculating features for two different bin widths.

We recently performed an extensive phantom study [16] (not part of this thesis), using the same physical texture phantom as has been used in previous studies [13, 14]. Among other things, we specifically investigated whether the bin width used for intensity discretization also has an effect on the test-retest stability of radiomic features. Results of this study indicated no significant effect of the choice of bin width on feature stability, which furthermore suggest radiomic features to be calculated for different bin widths to assess their potential complementary prognostic or predictive value.

Chapter 5 presented a study of the interoperability of two independent radiomics software implementations, namely the software developed as part of this thesis (addendum Software development) and in-house developed software from the University Hospital Zürich. This comparison was performed in the context of the development of a prognostic radiomic model for local tumor control in head and neck cancer, based on post-radiochemotherapy PET imaging. Both software implementations were each used to develop an independent local recurrence model, only considering features which were based on the same definition and available in both implementations. Both of these models were found to be prognostic for local tumor control in HNSCC and contained features that were highly reproducible between both software implementations. Subsequent validation resulted in similar model performance using either one of both software for feature calculation. However, univariable analysis revealed that the majority of features were not reproducible between both software implementations, indicating that interoperability of these different software is only limited.

Besides the aforementioned software, there are multiple other implementations of radiomics, which are either open-source, commercial or in-house developed [4, 17-19]. As discussed in **Chapter 9**, it is clear that this variation in software implementations, as well as nomenclature, mathematical definitions and methodology, makes reproducibility and validation of studies in radiomics a major challenge [1, 4, 17-20]. These differences have to be clarified, in order to facilitate interoperability of radiomics. As mentioned in **Chapter 5**, the image biomarker standardisation initiative (IBSI) is an independent international collaboration, which aims to address this challenge by standardization of image biomarkers [21]. The IBSI therefore sets out to provide a common nomenclature and definitions for image biomarkers, benchmarks for image processing and feature extraction, as well as reporting guidelines. Incorporating the results from this initiative in a radiomic ontology, by providing an extensive 'dictionary' [17, 22], could further facilitate multicenter studies. The IBSI also makes use of the digital radiomics phantom we introduced in **Chapter 9**, indicating the usefulness of such tools.

RADIOMICS IN LUNG AND HEAD AND NECK CANCER

The work presented in **Chapter 6** describes the development and validation of a prognostic radiomic signature, based on standard of care CT images, in 1019 non-small cell lung cancer and head and neck cancer patients. The results of this study provide a proof of concept that radiomics is able to decode a general prognostic phenotype, which independently validated both in lung and head and neck cancer. Illustrating the clinical relevance of the presented findings, the radiomic signature performed better in independent cohorts than TNM classification [23], which is routinely used in the clinic for treatment selection. Furthermore, radiogenomic analysis revealed that the signature, which is related to intra-tumor heterogeneity, correlates with underlying gene-expression patterns.

Chapter 7 presented further validation of the prognostic value of this radiomic signature, based on CT imaging from a cohort of 542 North American oropharyngeal squamous cell carcinoma patients. The signature was found to validate well in the entire validation cohort. Subsequently it was investigated how validation results are affected by the presence of CT image artifacts (e.g. streak artifacts), which are often present in CT images of head and neck cancer patients due to dental hardware [24]. Separate validation in the subset of patients with, as well as in the subset of patients without CT artifacts, revealed that the signature retained significant prognostic value, regardless of the presence of visible CT artifacts.

Even though the aforementioned radiomic signature was found to have prognostic value in both lung and head and neck cancer patients, we recently showed that CT based radiomic features exhibit both common as well as specific clustering for both disease

types [25] (not part of this thesis). These results indicate disease specific information of radiomic features, which should be exploited for the development of future models.

In **Chapter 8,** it was investigated whether human papilloma virus (HPV) status of OPSCC patients, determined by p16 immunohistochemistry, can be objectively identified by a quantitative radiomic approach. Previous exploratory studies were either based on small data sets without validation, or single institution data [26, 27]. The multicenter study presented in **Chapter 8** developed and validated a CT based radiomic signature to predict HPV status in 778 patients from four independent institutions. Although not intended to replace existing HPV tests, the results of this study provide a proof of concept that radiomics is able to derive molecular information from standard medical images. Another recent proof of concept study on genotype-phenotype interactions, furthermore illustrates the potential for developing noninvasive radiomic biomarkers for somatic mutations in NSCLC patients [28].

Nonetheless, as pointed out in **Chapter 9**, radiogenomics calls for a deeper understanding of the link between radiomic phenotypes and underlying biology [29-31]. Previous studies mainly reported correlations between radiomic features or signatures and gene expression patterns, such as described in **Chapter 6-7**, but causal relationships remain unclear and require future investigation [31]. To gain deeper understanding of potential causality, prospective pre-clinical studies could be proposed, such as to demonstrate whether genetic changes with phenotypic consequences influence radiomic features [32] (not part of this thesis).

FURTHER PERSPECTIVES AND CONCLUDING REMARKS

The majority of radiomic studies, including the work described in **Chapters 5-8** of this thesis only take into account a single time-point. Delta radiomics, as pointed out in **chapter 9**, is a concept where changes of features or signatures are evaluated, which was also illustrated in **Chapter 4**. Such a longitudinal approach could prove to be useful for early response assessment or treatment adaptation [33]. Longitudinal imaging is however scarce. Therefore, we recently proposed to perform delta radiomics on cone-beam CT (CBCT) images, which are routinely acquired over the course of radiotherapy treatment [34]. Even though CBCT images are generally of inferior quality compared to conventional treatment planning CT, we were able to show that a number of radiomic features, including the signature described in **Chapter 6-7**, are interchangeable between both imaging modalities using a two-step correction procedure [35] (not part of this thesis). Further research involved the development methodology to select CBCT derived radiomic features which are meaningful and informative in a longitudinal setting [36] (not part of this thesis). Future work will involve investigating the prognostic or predictive value of these features as well as establishing appropriate modeling techniques that allow for meaningful incorporation of longitudinal data.

By development of easy to perform and actionable biomarkers, radiomics furthermore has potential as patient stratification tool for novel therapeutic strategies. In Chapter 9, we provided the example of immunotherapy and hypothesized that tumors with high mutational loads will be more sensitive to immunotherapy and more heterogeneous on radiomic analysis. Application of radiomics for CDSS is therefore not only envisioned to result in less over- and under-treatment of patients, but as well allow for more effective, less costly clinical trial designs to bring new treatments to the clinic.

As discussed in Chapter 9, several challenges are associated with the workflow of radiomics, such as variability in derived features and segmentation of regions of interest. Furthermore, many methods for feature and model selection are available, where one should recognize that there is not a priori a single optimal machine learning method for any given problem [37-39]. Medical image analysis, including radiomics, is a constantly evolving field and deep learning is gaining increasing attention [40]. Deep learning has the potential to solve challenges posed by image segmentation, feature extraction and feature and model selection. The promise of deep learning methods is to discover new biomarkers without any human input, such as segmentations and handcrafted features. Instead, deep learning automatically discovers visual features that are suited to a specific task through a process of optimization, including features with varying levels of complexity [41, 42]. Although deep learning is a very promising development, a prominent challenge is the requirement of large amounts of data. Nonetheless, the added value of human defined radiomic features to deep learning should not be ignored and may be more practical and effective than substantially increasing the number of samples for training a deep learning model [43, 44]. As demonstrated in this thesis, radiomics has great potential in precision medicine, yet care should be taken with respect to interoperability. It is envisioned that global rapid learning healthcare networks [45, 46] will allow for continuous development, validation as well as revising of CDSS including radiomics. A key prerequisite to achieve this is for data to be made available according to the FAIR principle, which stands for Findability, Accessibility, Interoperability, and Reusability [47, 48]. Standardization initiatives, such as the IBSI, as well as international federated data sharing platforms employing ontologies for radiomic data [17, 22] (Chapter 9) all play an important role in this respect and are necessary to achieve widespread acceptance and application of radiomics in the clinic.

REFERENCES

[1] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278: 563-577.

[2] Aerts HJ. The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review. JAMA Oncol 2016;2: 1636-1642.

[3] Scrivener M, de Jong EEC, van Timmeren JE, Pieters T, Ghaye B, Geets X. Radiomics applied to lung cancer: a review. Translational Cancer Research 2016;5: 398-409.

[4] Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present... any future? Eur J Nucl Med Mol Imaging 2017;44: 151-165.

[5] Lee G, Lee HY, Ko ES, Jeong WK. Radiomics and imaging genomics in precision medicine. Precis Future Med 2017;1: 10-31.

[6] Larue RT, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. Br J Radiol 2017;90: 20160665.

[7] Parekh V, Jacobs MA. Radiomics: a new application from established techniques. Expert Review of Precision Medicine and Drug Development 2016;1: 207-226.

[8] Limkin EJ, Sun R, Dercle L, Zacharaki EI, Robert C, Reuzé S, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. Annals of Oncology 2017;28: 1191-1206.

[9] Scalco E, Rizzo G. Texture analysis of medical images for radiotherapy applications. The British Journal of Radiology 2017;90: 20160642.

[10] Bashir U, Siddique MM, McLean E, Goh V, Cook GJ. Imaging Heterogeneity in Lung Cancer: Techniques, Applications, and Challenges. AJR Am J Roentgenol 2016;207: 534-543.

[11] van Timmeren JE, Leijenaar RT, van Elmpt W, Wang J, Zhang Z, Dekker A, et al. Test–Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? Tomography 2016;2: 361-365.

[12] Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Scientific reports 2016;6: 23428.

[13] Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest Radiol 2015;50: 757-765.

[14] Shafiq-ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys 2017;44: 1050-1062.

[15] Larue RTHM, Van De Voorde L, van Timmeren JE, Leijenaar RTH, Berbee M, Sosef MN, et al. 4DCT imaging to assess radiomics feature stability: An investigation for thoracic cancers. Radiother Oncol 2017.

[16] Larue RTHM, van Timmeren JE, de Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, et al. Influence of grey level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncol In press.

[17] Kalpathy-Cramer J, Mamomov A, Zhao B, Lu L, Cherezov D, Napel S, et al. Radiomics of Lung Nodules: A Multi-Institutional Study of Robustness and Agreement of Quantitative Imaging Features. Tomography 2016;2: 430-437.

[18] Fang YH, Lin CY, Shih MJ, Wang HM, Ho TY, Liao CT, et al. Development and evaluation of an open-source software package "CGITA" for quantifying tumor heterogeneity with molecular images. Biomed Res Int 2014;2014: 248505.

[19] Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: An open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys 2015;42: 1341-1353.

[20] Yip SS, Aerts HJ. Applications and limitations of radiomics. Phys Med Biol 2016;61: R150-166.

[21] Zwanenburg A. EP-1677: Multicentre initiative for standardisation of image biomarkers. Radiotherapy and Oncology 2017;123: S914-S915.

[22] http://bioportal.bioontology.org/ontologies/RO/ (accessed on: 16-08-2017)

[23] Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol 2010;17: 1471-1474.

[24] Purohit BS, Ailianou A, Dulguerov N, Becker CD, Ratib O, Becker M. FDG-PET/CT pitfalls in oncological head and neck imaging. Insights into imaging 2014;5: 585-602.
[25] Parmar C, Leijenaar RT, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. Scientific reports 2015;5: 11044.
[26] Bogowicz M, Riesterer O, Ikenberg K, Stieb S, Moch H, Studer G, et al. CT radiomics predicts HPV status and local tumor control after definitive radiochemotherapy in head and neck squamous cell carcinoma. International Journal of Radiation Oncology*Biology*Physics 2017.
[27] Buch K, Fujita A, Li B, Kawashima Y, Qureshi MM, Sakai O. Using Texture Analysis to Determine Human Papillomavirus Status of Oropharyngeal Squamous Cell Carcinomas on CT. AJNR Am J Neuroradiol 2015;36: 1343-1348.
[28] Yip SS, Kim J, Coroller TP, Parmar C, Velazquez ER, Huynh E, et al. Associations Between Somatic Mutations and Metabolic Imaging Phenotypes in Non-Small Cell Lung Cancer. J Nucl Med 2017;58: 569-576.
[29] Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nat Biotechnol 2007;25: 675-680.
[30] Rosenstein BS, West CM, Bentzen SM, Alsner J, Andreassen CN, Azria D, et al. Radiogenomics: radiobiology enters the era of big data and team science. Int J Radiat Oncol Biol Phys 2014;89: 709-713.
[31] Incoronato M, Aiello M, Infante T, Cavaliere C, Grimaldi AM, Mirabelli P, et al. Radiogenomic Analysis of Oncological Data: A Technical Survey. Int J Mol Sci 2017;18.
[32] Panth KM, Leijenaar RT, Carvalho S, Lieuwes NG, Yaromina A, Dubois L, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. Radiother Oncol 2015;116: 462-466.
[33] Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Scientific reports 2017;7: 588.
[34] Fave X, Mackin D, Yang J, Zhang J, Fried D, Balter P, et al. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? Med Phys 2015;42: 6784-6797.
[35] van Timmeren JE, Leijenaar RTH, van Elmpt W, Reymen B, Oberije C, Monshouwer R, et al. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiother Oncol 2017.
[36] van Timmeren JE, Leijenaar RTH, van Elmpt W, Reymen B, Lambin P. Feature selection methodology for longitudinal cone-beam CT radiomics. Acta Oncol In press.
[37] Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, et al. Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology. Front Oncol 2016;6: 71.
[38] Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine Learning methods for Quantitative Radiomic Biomarkers. Scientific reports 2015;5: 13087.
[39] Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJ. Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. Front Oncol 2015;5: 272.
[40] Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng 2017;19: 221-248.
[41] Carneiro G, Oakden-Rayner L, Bradley AP, Nascimento J, Palmer L. Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography. In: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on: IEEE, 2017:130-134.
[42] Oakden-Rayner L, Carneiro G, Bessen T, Nascimento JC, Bradley AP, Palmer LJ. Precision Radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework. Scientific reports 2017;7: 1648.
[43] Paul R, Hawkins SH, Balagurunathan Y, Schabath MB, Gillies RJ, Hall LO, et al. Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma. Tomography 2016;2: 388-395.
[44] Kooi T, Litjens G, van Ginneken B, Gubern-Merida A, Sanchez CI, Mann R, et al. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis 2017;35: 303-312.

[45] Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CM, et al. 'Rapid Learning health care in oncology' - an approach towards decision support systems enabling customised radiotherapy'. Radiother Oncol 2013;109: 159-164.
[46] Lambin P, Zindler J, Vanneste B, van de Voorde L, Jacobs M, Eekers D, et al. Modern clinical research: How rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine. Acta Oncol 2015;54: 1289-1300.
[47] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3: 160018.
[48] Lustberg T, van Soest J, Jochems A, Deist T, van Wijk Y, Walsh S, et al. Big Data in radiation therapy: challenges and opportunities. Br J Radiol 2017;90: 20160689.

Software development

Ralph T.H. Leijenaar

INTRODUCTION

The workflow of radiomics (**Figure 1**) starts with the acquisition of medical images and segmentations of structures (i.e. regions of interest), such as the gross tumor volume defined for radiotherapy treatment planning purposes. Subsequently, large numbers of quantitative imaging features are extracted from the defined structures. These features can then be analyzed for their association with specific outcomes, in order to develop diagnostic, theragnostic, prognostic and predictive imaging biomarkers: so-called radiomic signatures.

Part of the work presented in this thesis consists of the development of software to enable the high-throughput extraction of radiomic features, as well as validated signatures [1], from medical images to facilitate further research in the field of radiomics. This chapter briefly describes the developed research software and its functionality.

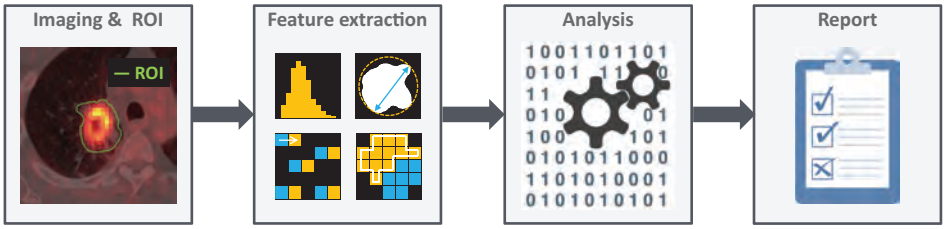


Figure 1 – The workflow of radiomics. The first step is the acquisition of medical images and segmentations of structures (i.e. regions of interest; ROI). Subsequently, large numbers of radiomic features are extracted from the defined structures. These features are then analyzed for their association with specific outcomes to develop radiomic signatures, based on which reports are generated for clinical decision support.

FUNCTIONALITY

The software is designed in Matlab (MathWorks, Natick, Massachusetts, USA) and provides three main functionalities: (1) data management, (2) radiomic workflow management, and (3) management of results.

Data

The software is able to process different imaging modalities, such as CT, PET and MR imaging, and associated structure segmentations. DICOM imaging data (CT, PET, or MR) and DICOM RTSTRUCT (containing segmentation contour data) are imported and converted and, once completed, the contents of the data can be (re)viewed. Radiotherapy structure sets (RTSTRUCT) usually contain delineations of multiple structures for treatment plan-

ning purposes and nomenclature of structures is not standardized. The software therefore provides an automated mapping tool to specify for each imported image dataset the appropriate structures to process (Figure 2).

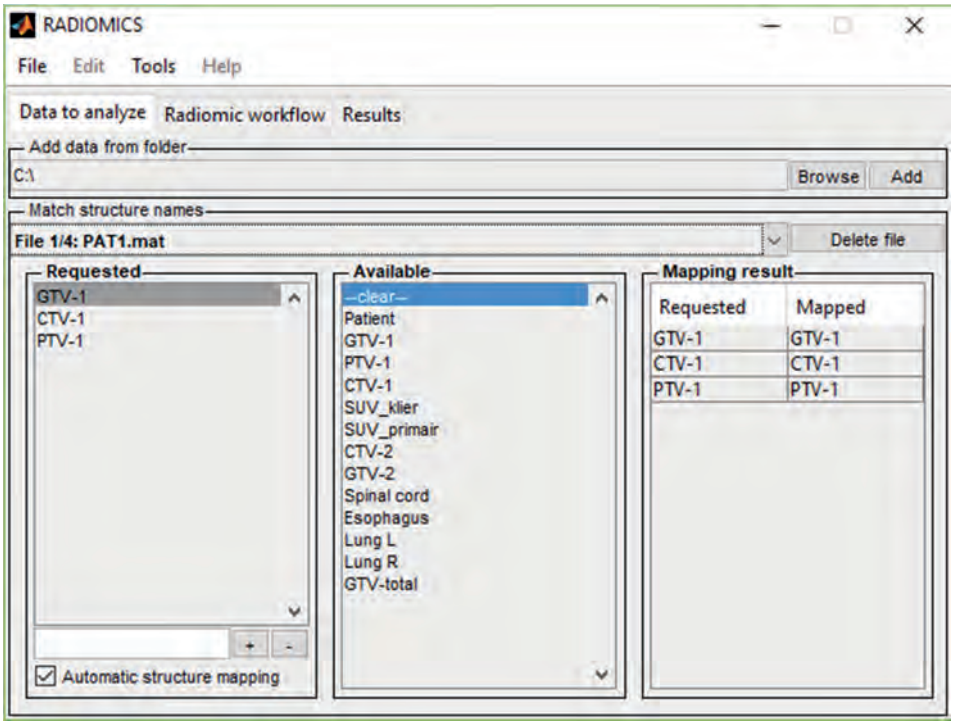


Figure 2 – Data management within the software. DICOM imaging data (CT, PET, and MR) and DICOM RTSTRUCT (structure contours) are imported. Once imported, the imaging and contour data can be (re)viewed. In this example, we have imported DICOM data of four patients. The software will, for each image dataset, display the available structures in the panel "Available". To select appropriate structure(s) to be analyzed, names of structures (e.g. GTV-1) to be processed can be inserted in the text field in the bottom of the panel "Requested". Automatic structure mapping is provided and, if this option is selected, the software will automatically map (i.e. link) the requested structure names to structures present (using e.g. regular expressions and partial name matching), for all imported image data. Structure mappings are displayed in the panel "Mapping result" and can be changed as desired.

Radiomic workflow

A radiomic workflow can be easily constructed for these images and structures. Such a workflow contains modules which are defined for each specific imaging modality (i.e. CT, PET, and MR). As many modules as desired, all with their own specific settings (e.g. selection of feature groups, parameters for feature extraction [2] and image preprocessing),

can be included in a workflow (**Figure 3**). The software allows to process numerous medical images and structures by creating only a single radiomic workflow (i.e. batch processing), which makes it a true high-throughput approach.

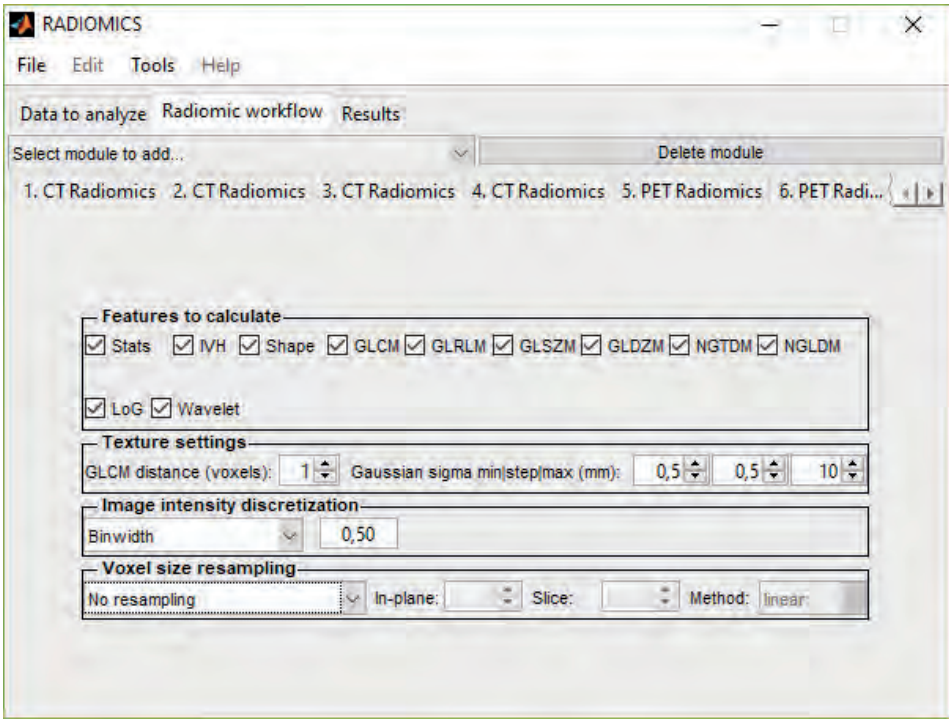


Figure 3 – Workflow management within the software. Workflow modules can be added for each specific modality (CT, PET, and MR). As many modules as desired, all with their own specific settings, can be added to the workflow.

Results

The results for the complete workflow are generated by 'a click of a button'. The output, containing all relevant feature and signature data, is generated in spreadsheet format and can be exported in either XLSX or CSV format for further processing or analysis (**Figure 4**).

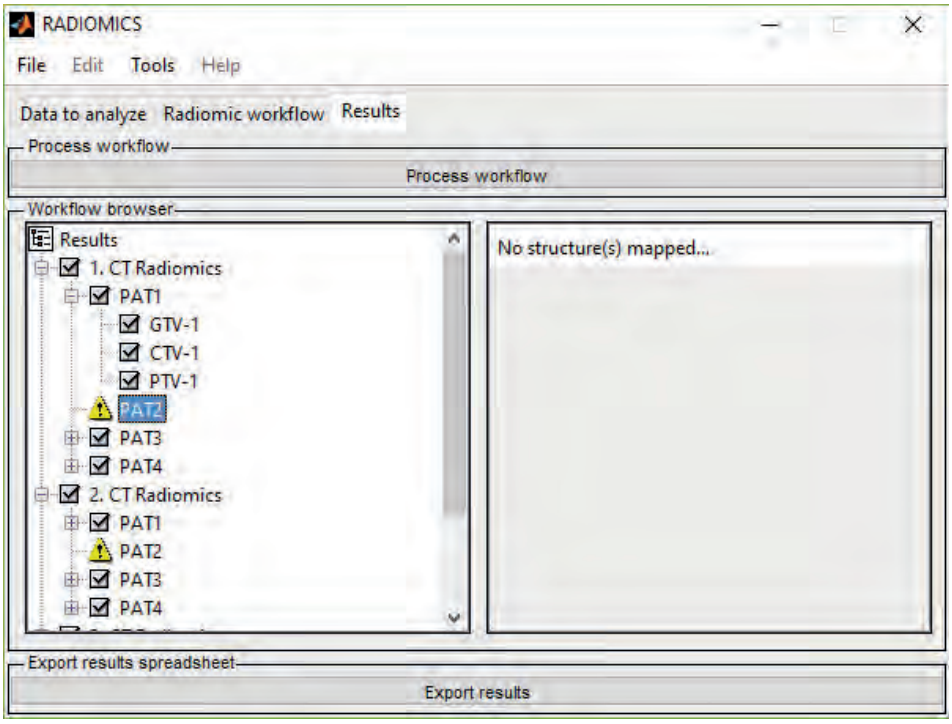


Figure 4 – Results management within the software. Here, the specified radiomic workflow is executed, generating results for each workflow module and specified structure for each included image data set. Once the processing of the workflow is finished, the output, containing all relevant feature and signature data, is exported in either XLSX or CSV format for further processing or analysis. A green checkbox is shown next to complete and processed data. A yellow exclamation mark indicates either missing data (e.g. when there is no structure mapped for a specific image set), or structures that have not (yet) been processed.

IMPLEMENTED RADIOMIC FEATURES AND SIGNATURE

The software provides a comprehensive, three-dimensional quantification of any provided region of interest, by computation of features of the following groups: (I) first-order, (II) geometric, (III) texture, and (IV) filtered features.

First-order statistics (group I) describe the distribution of image voxel intensity values within the region of interest. Geometric features (group II) describe the shape and size of the region of interest. Textural features (group III) are derived from grey-level co-occurrence (GLCM) [3], grey-level distance-zone (GLDZM) [4], grey-level run-length (GLRLM) [5], grey-level size-zone (GLSZM) [6, 7], neighbouring grey-level dependence (NGLDM) [8], and neighbourhood grey-tone difference (NGTDM) [9] matrices. Filtered features (group IV) are features calculated after wavelet decomposition and after application of

Laplacian of Gaussian filters. An undecimated three-dimensional wavelet transform is applied to each image, which decomposes the original image into 8 decompositions. Statistics (group I) and textural (group III) features are calculated for each decomposition [1]. By applying a Laplacian of Gaussian filter, textural properties representing features of different degrees of coarseness can be calculated. The degree of coarseness (fine to coarse) is determined by the Gaussian radius parameter σ. Each value of σ provides a filtered image. First-order gray-level statistics (group I) are determined for each filtered image, as well as for only the positive part of each filtered image [10]. Detailed feature descriptions as well as mathematical definitions can be found in literature [10-13].

Furthermore, a validated prognostic radiomic signature can be calculated, which is based on standard CT imaging of over 1000 non-small cell lung cancer and head and neck cancer patients. This signature is further described in Chapter 6 of this thesis [1].

REFERENCES

[1] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5: 4006.

[2] Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Scientific reports 2015;5: 11075.

[3] Haralick RM, Shanmugam K, Dinstein I. Textural Features of Image Classification. IEEE T Syst Man Cyb 1973;SMC-3: 610-621.

[4] Thibault G, Angulo J, Meyer F. Advanced Statistical Matrices for Texture Characterization: Application to Cell Classification. IEEE Transactions on Biomedical Engineering 2014;61: 630-637.

[5] Galloway M. Texture analysis using gray level run lengths. Comput Vision Graph 1975;4: 172-179.

[6] Thibault GF, B; Navarro, C; Pereira, S. Texture indexes and gray level size zone matrix: application to cell nuclei classification. Pattern Recognition Inf Process. 2009: 140-145.

[7] Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. J Nucl Med 2012;53: 693-700.

[8] Sun C, Wee WG. Neighboring gray level dependence matrix for texture classification. Computer Vision, Graphics, and Image Processing 1983;23: 341-352.

[9] Amadasun M, King R. Textural features corresponding to textural properties. Systems, Man and Cybernetics, IEEE Transactions on 1989;19: 1264-1274.

[10] van Timmeren JE, Leijenaar RTH, van Elmpt W, Reymen B, Oberije C, Monshouwer R, et al. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiother Oncol 2017.

[11] Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RT, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 2015;114: 345-350.

[12] Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol 2013;52: 1391-1397.

[13] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol In press.

Valorisation

Ralph T.H. Leijenaar

INTRODUCTION

Increasing evidence of inter- and intra-patient heterogeneity and an ever increasing number of novel treatment strategies, mean there is a major need for the identification of non-invasive and easy to repeat biomarkers to be incorporated in clinical decision support systems (CDSS) to guide precision medicine.

Radiomics concerns with the high-throughput mining of large amounts of quantitative features from standard medical images, for knowledge extraction. As described in this thesis, radiomics has great potential to improve CDSS, by providing complementary and interchangeable information alongside other sources, such as demographics, pathology, genomics and proteomics.

The research carried out at Maastricht Radiation Oncology (MAASTRO) clinic and other institutes worldwide, has resulted in a large number of publications, which has shown a rapid increase during the last few years. Many possible applications for radiomics are extensively discussed in recent literature, which shows its great valorisation potential. This undeniable potential has led to a concrete business plan, based on which the spin-off company Oncoradiomics has been established in Liège, Belgium, in 2016.

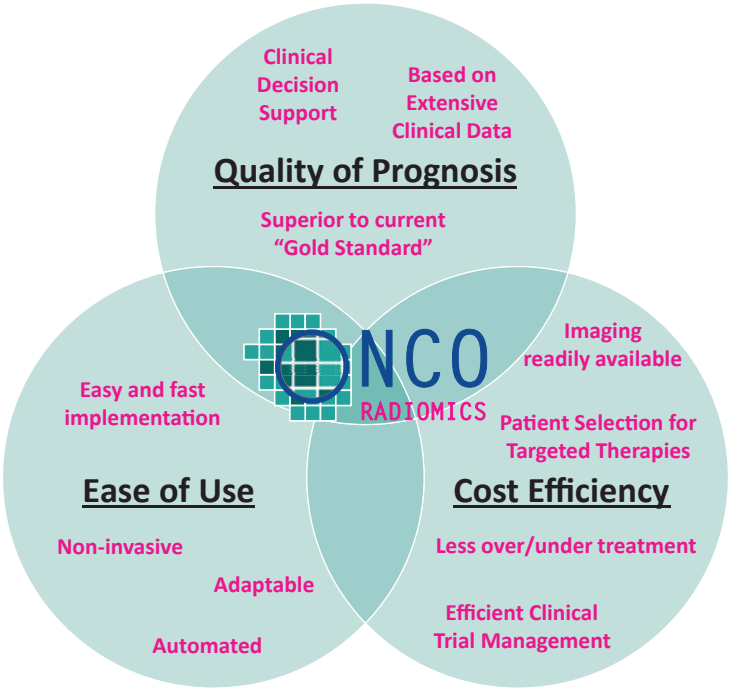


Figure 1 – Oncoradiomics aims to develop easy to use products and services providing higher quality of care for cancer patients while increasing the cost-effectiveness of treatment.

ONCORADIOMICS

Oncoradiomics (www.oncoradiomics.com) is a life science company and its intellectual property originates from MAASTRO clinic. Oncoradiomics aims to develop easy to use products and services providing higher quality of care (e.g. prognosis) for cancer patients while increasing the cost-effectiveness of treatment (Figure 1).

RADIOMIX

Based on the software developed for this thesis (addendum Software development), Oncoradiomics has developed RADIOMIX. RADIOMIX is a clinical grade CE marked software solution, enabling the extraction of quantitative image features and radiomic signatures from standard medical images to guide precision oncology. It includes the validated prognostic signature described in Chapter 6-7 of this thesis. Different versions of this software will be made available to tailor different segments of the market.

First, it is made available as a plug-in, which makes it possible to easily integrate radiomics into existing software packages of large established market players in the fields of radiology, radiation therapy, PACS and other healthcare IT software. One example is the recent integration of radiomics into Aquilab's ARTIVIEW software, which has been presented at ESTRO 35 (29 April - 3 May 2016, Turin, Italy) and ESTRO 36 (5-9 May 2017, Vienna, Austria).

RADIOMIX will also be available as a cloud based software as a service. As such, it will provide means for patient stratification and response assessment using a secure online platform. This also makes it an ideal platform for clinical trials to investigate the effectiveness of novel therapeutic strategies or drugs.

Since the software is development in a modular way, future radiomic features and signatures can be easily added to the system to be made available to end users.

Currently, several renowned cancer centres are using a research version of RADIOMIX. Development of novel radiomic signatures will initially be investigated in close collaboration with these select centres. This research will result in presentations at key conferences, as well publication of peer-reviewed scientific articles.

DISTRIM

As discussed in this thesis, radiomics is fundamentally dependent upon the availability of data for development and validation. Oncoradiomics therefore also develops a commercial solution for distributed learning, DISTRIM. DISTRIM aims to enable a globally scalable and encrypted network, which will allow for (1) the development and validation of new radiomic signatures, and (2) continuous refinement and updating of existing signatures, to increase prognostic, predictive or diagnostic power.

Dankwoord

DANKWOORD

Eindelijk is het zover! Na jaren onderzoek, waar ik met veel plezier op terugkijk, is mijn proefschrift voltooid. Graag wil ik iedereen bedanken die direct of indirect een bijdrage heeft geleverd aan mijn werk.

Ten eerste wil ik mijn promotor prof. dr. Philippe Lambin en mijn copromotoren dr. ir. Wouter van Elmpt en dr. Frank Hoebers bedanken voor al hun hulp en hun waardevolle bijdrage aan dit proefschrift. Philippe, u bent een echte bron van inspiratie en motivatie. Niet alleen op wetenschappelijk gebied, maar ook op het gebied van academic entrepreneurship. Mijn dank daarvoor. Wouter, bij jou kon ik altijd terecht voor vragen en discussie. Je gaat geen onderwerp uit de weg en zoekt altijd mee naar oplossingen. Jouw input heeft me geholpen het beste in mezelf naar boven te halen. Frank, hartelijk dank voor onze discussies en je hulp bij de hoofdhals projecten. Met name bedankt voor je ondersteuning betreffende onze internationale samenwerking met Toronto, een essentieel onderdeel van dit proefschrift.

Daarnaast wil ik de leden van de beoordelingscommissie (prof. dr. ir. W. Backes, prof. dr. F. Ramaekers, prof. dr. M. Dumontier, prof. dr. W. Niessen en prof. dr. M. van den Brekel) bedanken voor het kritisch lezen en beoordelen van mijn thesis.

Graag wil ik ook alle coauteurs bedanken voor het totstandkomen van de manuscripten die deel uitmaken van dit proefschrift.

Sara, you were the other half of the team and my neighbour for many years. It really was a pleasure working with you. I do hope your nerves didn't suffer permanent damage! Stefan, CEO^2, we hebben samen menig kopje koffie weggetikt de afgelopen jaren. Er was dan ook altijd wel iets om over te praten. Dankjewel voor alle mooie momenten. Veel succes met ondernemen en vooral met de laatste loodjes voor jouw eigen PhD. Evelyn, Aniek, jullie zijn geweldige overburen. Samen met Ruben, Janita en Jurgen is het altijd gezellig daar in de hoek van de research ruimte. Bedankt dat jullie mijn (niet altijd even relevante) verhalen willen aanhoren en dat jullie altijd bereid zijn om je mening te geven. Karen, altijd vrolijk en ook altijd in voor een goed gesprek. Bedankt daarvoor. Unfortunate 'Shane' of events, thank you for your always amusing (and amazing) stories. Seán, buddy, you are a great guy and an awesome colleague. Thank you as well for your invaluable input. Marta, thank you for the great collaboration and many fruitful discussions we had. Hugo en Emmanuel, ook jullie mogen natuurlijk niet ontbreken. Heel erg bedankt voor alle support aan het begin van mijn avontuur. Rianne, dankjewel voor alle (administratieve) ondersteuning.

Special thanks also go to all my other fellow researchers (in no particular order): Henry, Arthur, Turkey, Sebastian, Simon, Abdalla, Kranthi, Brent, Adriana, Lucas, Davide, Alberto, Ana, Cecile, Daniela R., Esther, Frank, Gabriel, Giacomo, Inge, Isabel, Jean, Johan, Jose, Jurgen, Leonard, Lotte, Mark, Murillo, Relinde, Scott, Stefan, Tim, Timo, Yvonka, Nicolle, Marco, Skadi, Celine, Daniela T., Patrick Granton, Guillaume, Ruud, Georgi, Francesco, Fiere, Georgy, Chintan, Patrick Grossman, Thibaud, Ala, Cary, Maud, Jaap, Bianca,

Casper, Maikel, Hoda, Tessa, Abir, Pouya, Bregtje, Pedro, Raghu, Mathieu, Matilde. I have crossed paths with many people and it is hard listing everybody, but it was a pleasure getting to know all of you and I enjoyed all the (fun) discussions, (walking) lunches, social events, trips (especially Oktoberfest and skiing) and conference visits we shared. Thank you!

Verder wil ik graag mijn familie en vrienden bedanken. Het was voor jullie wellicht niet altijd duidelijk wat mijn werk nu precies inhield, maar mede dankzij jullie is dit proefschrift tot stand gekomen. Dick(y) en Wesley, paranimfen, van al mijn vrienden zijn jullie er het langst. We kennen elkaar nu al zo'n 30 jaar en jullie staan altijd voor me klaar. Ik hoop dat we nog veel samen mogen meemaken! Tante Tiny, ome Jacques, bedankt dat jullie er altijd voor me zijn geweest. Céline, mijn grote kleine zusje, I did it! 😊

Tenslotte gaat mijn grootste dank uit naar mijn ouders. Papa en mama, jullie zijn altijd in me blijven geloven en hebben me waar mogelijk ondersteund. Pap, zonder jouw hulp zou mijn woning nooit zo mooi zijn geworden en datzelfde geldt ook voor dit boekje. Bedankt voor het mooie schilderij voor op de omslag, een echte "Huub".

Curriculum Vitae

CURRICULUM VITAE

Ralph Leijenaar was born on the 5th of July 1984 in Heerlen in the Netherlands. After high school at the st.-Janscollege in Hoensbroek, he studied Biomedical Engineering at the Eindhoven University of Technology (TU/e). During the initial years of his studies, he developed special interest for the fields of computer science and image analysis. Throughout his master studies, he was able to combine these interests in an internship at Maastricht Radiation Oncology (MAASTRO) clinic. During this internship he investigated the use of advanced imaging features, extracted from CT images before and during treatment, for the prediction of survival in non-small cell lung cancer (NSCLC). His master thesis project was another joint project between MAASTRO clinic and the TU/e. Here he investigated the relationship between tumor FDG-PET uptake and advanced CT imaging information in NSCLC. After completion of his graduation work, he obtained his Master of Science degree in 2011. His master project was one of the pioneering steps in the field of radiomics. He continued his scientific career at MAASTRO clinic as a PhD student under the supervision of Prof. dr. Philippe Lambin. In 2017, he completed his PhD thesis in the multidisciplinary field of radiomics. As an academic entrepreneur, he currently is the Chief Technology Officer of OncoRadiomics and one of its founders. The development of a commercial clinical grade CE marked radiomics solution is his principal responsibility.

Awards

28-01-2016 **GROW Valorisation Award.** Awarded by GROW, the School for Oncology and Developmental Biology at the Maastricht University Medical Centre (MUMC+), The Netherlands.

19-04-2016 **NRS Science Award.** Awarded by the Netherlands Respiratory Society to Philippe Lambin, Anne-Marie Dingemans and Ralph Leijenaar for the work on “Radiomics in lung cancer”.

Patents

Image analysis method supporting illness development prediction for a neoplasm in a human or animal body. Application No.: PCT/NL2014/050728, filing date 17-10-2014, WO 2016060557 A1.

List of publications

LIST OF PUBLICATIONS

1. Lambin P, **Leijenaar RTH**. Radiomics: meer informatie uit medisch beeldmateriaal. Oncologie Up-to-date. 2012;3(2).

2. Lambin P, Rios-Velazquez E, **Leijenaar RTH**, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48(4):441-6.

3. Carvalho S, **Leijenaar RTH**, Velazquez ER, Oberije C, Parmar C, van Elmpt W, Reymen B, Troost EG, Oellers M, Dekker A, Gillies R, Aerts HJ, Lambin P. Prognostic value of metabolic metrics extracted from baseline positron emission tomography images in non-small cell lung cancer. Acta Oncol. 2013;52(7):1398-404.

4. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CM, Carvalho S, **Leijenaar RTH**, Nalbantov G, Oberije C, Scott Marshall M, Hoebers F, Troost EG, van Stiphout RG, van Elmpt W, van der Weijden T, Boersma L, Valentini V, Dekker A. 'Rapid Learning health care in oncology' - an approach towards decision support systems enabling customised radiotherapy'. Radiother Oncol. 2013;109(1):159-64.

5. **Leijenaar RTH**, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, Hoekstra CJ, Boellaard R, Dekker AL, Gillies RJ, Aerts HJ, Lambin P. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. 2013;52(7):1391-7.

6. Aerts HJ, Velazquez ER, **Leijenaar RTH**, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, Hoebers F, Rietbergen MM, Leemans CR, Dekker A, Quackenbush J, Gillies RJ, Lambin P. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.

7. Hoeben BA, Starmans MH, **Leijenaar RTH**, Dubois LJ, van der Kogel AJ, Kaanders JH, Boutros PC, Lambin P, Bussink J. Systematic analysis of 18F-FDG PET and metabolism, proliferation and hypoxia markers for classification of head and neck tumors. BMC cancer. 2014;14:130.

8. Parmar C, Rios Velazquez E, **Leijenaar RTH**, Jermoumi M, Carvalho S, Mak RH, Mitra S, Shankar BU, Kikinis R, Haibe-Kains B, Lambin P, Aerts HJ. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One. 2014;9(7):e102107.

9. Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, **Leijenaar RTH**, Hermann G, Lambin P, Haibe-Kains B, Mak RH, Aerts HJ. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol. 2015;114(3):345-50.

10. Lambin P, Zindler J, Vanneste B, van de Voorde L, Jacobs M, Eekers D, Peerlings J, Reymen B, Larue RT, Deist TM, de Jong EE, Even AJ, Berlanga AJ, Roelofs E, Cheng Q, Carvalho S, **Leijenaar RTH**, Zegers CM, van Limbergen E, Berbee M, van Elmpt W, Oberije C, Houben R, Dekker A, Boersma L, Verhaegen F, Bosmans G, Hoebers F, Smits K, Walsh S. Modern clinical research: How rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine. Acta Oncol. 2015;54(9):1289-300.

11. **Leijenaar RTH**, Carvalho S, Hoebers FJ, Aerts HJ, van Elmpt WJ, Huang SH, Chan B, Waldron JN, O'Sullivan B, Lambin P. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta Oncol. 2015;54(9):1423-9.

12. **Leijenaar RTH**, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, Aerts HJ, Gillies RJ, Lambin P. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Scientific reports. 2015;5:11075.

13. Panth KM, **Leijenaar RTH**, Carvalho S, Lieuwes NG, Yaromina A, Dubois L, Lambin P. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. Radiother Oncol. 2015;116(3):462-6.

14. Parmar C, **Leijenaar RTH**, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, Rietbergen MM, Haibe-Kains B, Lambin P, Aerts HJ. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. Scientific reports. 2015;5:11044.

15. Trani D, Reniers B, Persoon L, Podesta M, Nalbantov G, **Leijenaar RTH**, Granzier M, Yaromina A, Dubois L, Verhaegen F. What Level of Accuracy Is Achievable for Preclinical Dose Painting Studies on a Clinical Irradiation Platform? Radiation research. 2015;183(5):501-10.

16. van Timmeren JE, **Leijenaar RTH**, van Elmpt W, Wang J, Zhang Z, Dekker A, Lambin P. Test–Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? Tomography. 2016;2(4):361-5.

17. Bogowicz M, **Leijenaar RTH**, Tanadini-Lang S, Riesterer O, Pruschy M, Studer G, Unkelbach J, Guckenberger M, Konukoglu E, Lambin P. Post-radiochemotherapy PET radiomics in head and neck cancer - The influence of radiomics implementation on the reproducibility of local control tumor models. Radiother Oncol. 2017.

18. de Jong EE, van Elmpt W, **Leijenaar RTH**, Hoekstra OS, Groen HJ, Smit EF, Boellaard R, van der Noort V, Troost EG, Lambin P, Dingemans AC. [18F]FDG PET/CT-based response assessment of stage IV non-small cell lung cancer treated with paclitaxel-carboplatin-bevacizumab with or without nitroglycerin patches. Eur J Nucl Med Mol Imaging. 2017;44(1):8-16.

19. Grossmann P, Stringfield O, El-Hachem N, Bui MM, Rios Velazquez E, Parmar C, **Leijenaar RTH**, Haibe-Kains B, Lambin P, Gillies RJ, Aerts HJ. Defining the biological basis of radiomic phenotypes in lung cancer. Elife. 2017;6:e23421.

20. Lambin P, **Leijenaar RTH**, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, Sanduleanu S, Larue RTHM, Even AJG, Jochems A, van Wijk Y, Woodruff HC, van Soest J, Lustberg T, Roelofs E, van Elmpt W, Dekker A, Mottaghy FM, Wildberger JE, Walsh S. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017.

21. Lambin P, Zindler J, Vanneste BG, De Voorde LV, Eekers D, Compter I, Panth KM, Peerlings J, Larue RT, Deist TM, Jochems A, Lustberg T, van Soest J, de Jong EE, Even AJ, Reymen B, Rekers N, van Gisbergen M, Roelofs E, Carvalho S, **Leijenaar RTH**, Zegers CM, Jacobs M, van Timmeren J, Brouwers P, Lal JA, Dubois L, Yaromina A, Van Limbergen EJ, Berbee M, van Elmpt W, Oberije C, Ramaekers B, Dekker A, Boersma LJ, Hoebers F, Smits KM, Berlanga AJ, Walsh S. Decision support systems for personalized and participative radiation oncology. Adv Drug Deliv Rev. 2017;109:131-53.

22. Larue RTHM, Van De Voorde L, van Timmeren JE, **Leijenaar RTH**, Berbee M, Sosef MN, Schreurs WMJ, van Elmpt W, Lambin P. 4DCT imaging to assess radiomics feature stability: An investigation for thoracic cancers. Radiother Oncol. 2017;125(1):147-53.

23. Larue RTHM, van Timmeren JE, de Jong EEC, Feliciani G, **Leijenaar RTH**, Schreurs WMJ, Sosef MN, Raat F, van der Zande FHR, Das M, van Elmpt W, Lambin P. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncol. 2017:1-10.

24. **Leijenaar RTH**, de Jong EEC, Larue RTHM, van Timmeren JE, Lambin P. Radiomics: de toekomst in medische beeldvorming. Ned Tijdsch Oncol. 2017;14:82-9.

25. Mihl C, Maas M, Turek J, Seehofnerova A, **Leijenaar RTH**, Kok M, Lobbes MB, Wildberger JE, Das M. Contrast Media Administration in Coronary Computed Tomography Angiography - A Systematic Review. Rofo. 2017;189(4):312-25.

26. Ou D, Blanchard P, Rosellini S, Levy A, Nguyen F, **Leijenaar RTH**, Garberis I, Gorphe P, Bidault F, Ferte C, Robert C, Casiraghi O, Scoazec JY, Lambin P, Temam S, Deutsch E, Tao Y. Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to Human Papillomavirus status. Oral oncology. 2017;71:150-5.

27. van Timmeren JE, **Leijenaar RTH**, van Elmpt W, Reymen B, Lambin P. Feature selection methodology for longitudinal cone-beam CT radiomics. Acta Oncol. 2017:1-7.

28. van Timmeren JE, **Leijenaar RTH**, van Elmpt W, Reymen B, Oberije C, Monshouwer R, Bussink J, Brink C, Hansen O, Lambin P. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiother Oncol. 2017;123(3):363-9.

29. Zindler JD, Jochems A, Lagerwaard FJ, Beumer R, Troost EGC, Eekers DBP, Compter I, van der Toorn PP, Essers M, Oei B, Hurkmans CW, Bruynzeel AME, Bosmans G, Swinnen A, **Leijenaar RTH**, Lambin P. Individualized early death and long-term survival prediction after stereotactic radiosurgery for brain metastases of non-small cell lung cancer: Two externally validated nomograms. Radiother Oncol. 2017;123(2):189-94.

30. Carvalho S, **Leijenaar RTH**, Troost EGC, van Timmeren JE, Oberije C, van Elmpt W, de Geus-Oei LF, Bussink J, Lambin P. FDG-PET-Radiomics of metastatic lymph nodes and primary tumor in NSCLC – a prospective externally validated study. Submitted work.

31. Larue RTHM, Klaassen R, Jochems A, **Leijenaar RTH**, Hulshof MCCM, van Berge Henegouwen MI, Schreurs WMJ, Sosef MN, van Elmpt W, van Laarhoven HW, Lambin P. Pre-treatment CT radiomics to predict 3-year overall survival following chemoradiotherapy of oesophageal cancer. Submitted work.

32. **Leijenaar RTH**, Bogowicz M, Jochems A, Hoebers FJP, Wesseling FWR, Huang SH, Chan B, Waldron J, O'Sullivan B, Rietveld D, Leemans CR, Brakenhoff RH, Riesterer O, Tanadini-Lang S, Guckenberger M, Lambin P. Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study. Submitted work.

33. Verduin M, Compter I, Steijvers D, Jacobi-Postma AA, Eekers DBP, Anten M, Ackermans L, ter Laan M, **Leijenaar RTH**, van de Weijer T, Tjan-Heijnen VC, Hoeben A, Vooijs M. Non-invasive glioblastoma testing: multimodal approach to monitoring and predicting treatment response. Submitted work.